

Transcriptome and genome sequencing uncovers functional variation in human populations

Tuuli Lappalainen et al.

Supplementary Material

Index

Supplementary Figures

Figure S1: Study design	4
Figure S2: Analysis of regulatory genetic variation	5
Figure S3. Genotype data quality control	6
Figure S4. Read and gene count distributions.....	7
Figure S5. miRNA quality control.....	7
Figure S6. Sample clustering.....	8
Figure S7. PEER covariate analysis	9
Figure S8. Sample clustering after normalization.....	10
Figure S9. Replicate correlation after normalization	11
Figure S10. Gene discovery in a population sample	11
Figure S11. Transcript quantification statistics.....	12
Figure S12. Transcript variation between population pairs	13
Figure S13. Tandem variation in splicing.....	14
Figure S14. Chimeric transcripts.....	15
Figure S15. miRNA quantification statistics.....	16
Figure S16. Coexpression of exons of the same gene	16
Figure S17. eQTL-trQTL sharing.....	17
Figure S18. Transcribed repeat eQTLs	17
Figure S19. Trans-effects of mirQTLs.....	18
Figure S20. RNA editing QTLs	19
Figure S21. Indel enrichment in eQTL variants	20
Figure S22. Functional annotation of eQTLs.....	21
Figure S23. Functional annotation of trQTLs.....	22
Figure S24. Causal eQTL variants	24
Figure S25. Overlap of eQTLs with Omni 2.5M SNPs	24
Figure S26. GWAS signal of eQTLs.....	25
Figure S27. Causal GWAS variants prediction.....	26
Figure S28. Quality control of ASE data.....	27
Figure S29. Population variation in ASE	28
Figure S30. Population variation across ASE frequency spectrum	29
Figure S31. Frequency spectrum of allele-specific transcript structure.....	30
Figure S32. Nonsense-mediated decay.....	30
Figure S33. Splice scores.....	31
Figure S34. The Geuvadis Data Browser.....	32

Supplementary Methods

Study design	33
RNA-sequencing data production.....	33
Cell line processing.....	33
RNA extraction	33
RNA sequencing.....	34
Raw data processing.....	34
Genotype data.....	34
Variant annotation.....	35
Imputation.....	35
Quality control.....	35
mRNA read mapping	35
mRNA quantifications	36
Exons and genes	36
Transcripts, splice junctions, and introns	36
Exon inclusion	37
Transcribed repeats.....	37
small RNA (sRNA) data processing.....	37
Improved miRNA gene annotations.....	37
sRNA read data processing.....	38
sRNA mapping and quantification	38
RNA-seq quality control.....	38
Outlier detection.....	38
Sample swap and contamination analysis	39
miRNA data quality control	39
Normalization of quantifications	39
Quantitative versus qualitative variation.....	40
Differentially transcribed genes.....	40
Chimeric transcripts	41
RNA editing.....	41
miRNA effects on the transcriptome	42
miRNA family and target definition	42
Integrated analysis of miRNA and mRNA expression	42
Trans-eQTL effects of cis-mirQTLs.....	43
Transcriptome QTL analysis.....	44
Transcriptome QTL mapping with linear regression.....	44
Transcript ratio QTL effects	44
Independence of QTLs	44
Null variant distribution	44
Causal regulatory variant estimation	45
GWAS overlap of eQTLs.....	45
Allele-specific analysis.....	46
Allele-Specific Expression (ASE) analysis.....	46
Allele-Specific Transcript Structure (ASTS) analysis	46
Loss-of-function analysis.....	47
Nonsense-mediated decay	47
Splice scores	47
The Geuvadis Data Browser	47
References to Supplementary Methods	48

Supplementary Tables

Table S1. Samples.....	52
Table S2. Variant annotations.....	53
Table S3. Quantifications.....	54
Table S4. Chimeric transcripts (legend).....	54
Table S5. Associated miRNA-mRNA pairs (legend)	54
Table S6. Predicted causal GWAS variants (legend).....	55

Supplementary Figures

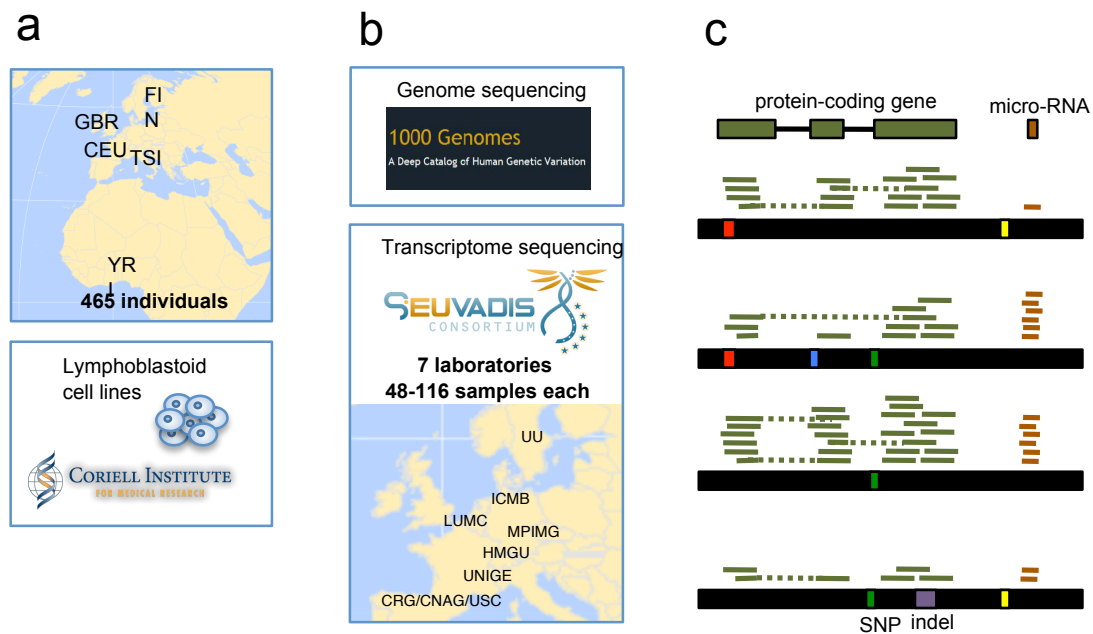


Figure S1: Study design

An illustration of the study design shows the studied populations and samples (a) from which the 1000 Genomes Consortium created genome sequencing and genotype data, and we sequenced mRNA and small RNA in seven European laboratories (b), with the final data set consisting of genotype and RNA-sequencing data from 462 and 452 individuals for mRNA and small RNA, respectively (c).

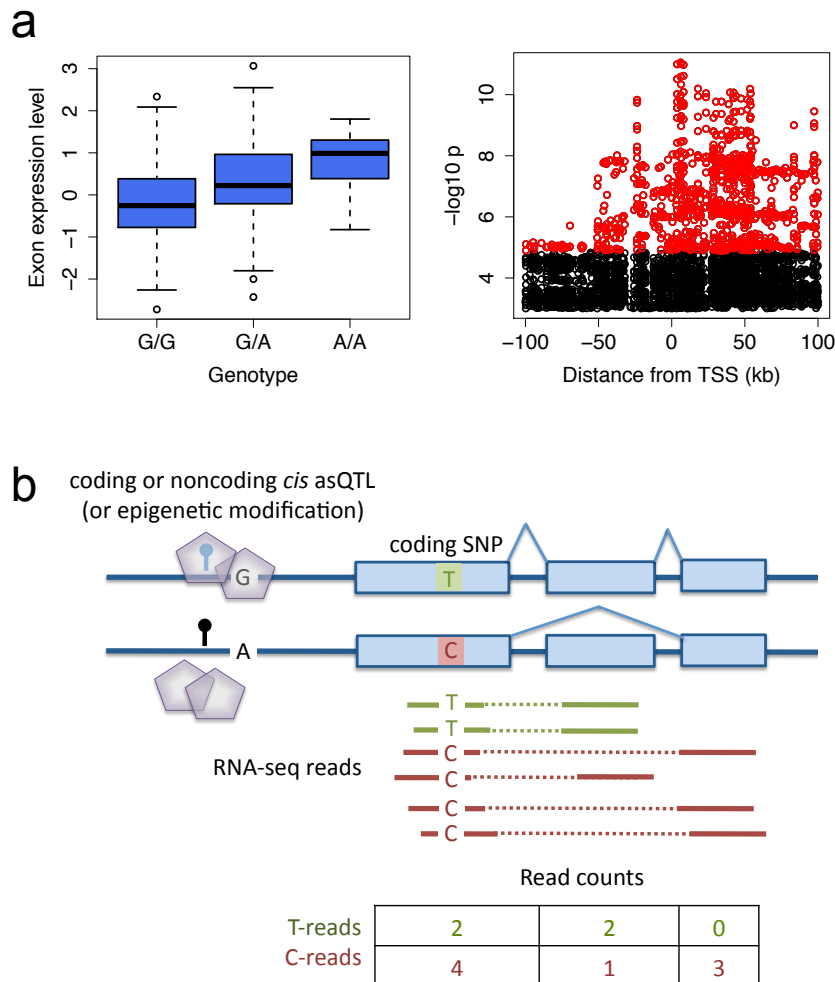


Figure S2: Analysis of regulatory genetic variation

Analytical approaches for studying genetic effects on transcription from RNA-sequencing data: a) transcriptome QTL analysis, where the aim is to find genetic variants that associate to a transcriptome quantitative trait (such as gene expression level) in a population sample. First, association between genotypes and transcriptome quantitative traits is calculated for all variant – transcript feature pairs usually in a genomic window (left panel), and the resulting p-values can be plotted as a landscape of associations surrounding the gene (right panel; red dots are genome-wide significant associations). (b) illustrates allele-specific transcription analysis that aims to identify differences in transcription between the two haplotypes of an individual, either in the ratio of the two alleles (allele-specific expression analysis or ASE) or in the exonic distribution of reads (allele-specific transcript structure or ASTS).

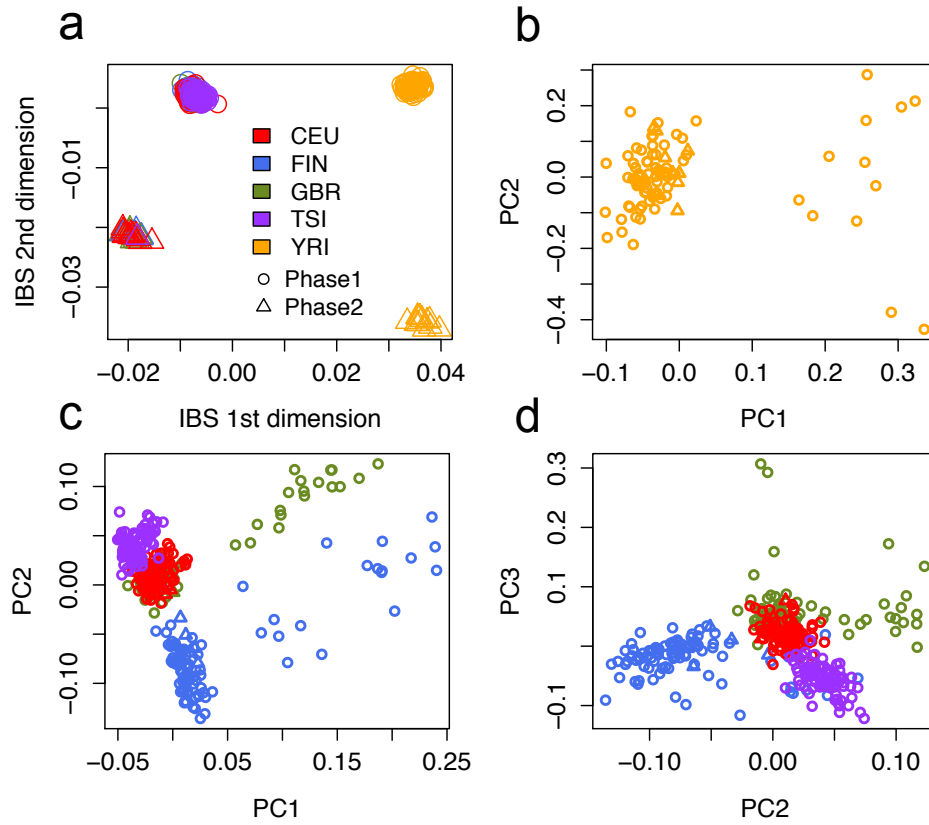


Figure S3. Genotype data quality control

Multidimensional scaling plot of identity-by-state matrix of all the samples shows a clear clustering not only by continent, but also by whether the sample had full genome data or was imputed (a). Principal component analysis within Yoruba (b) and within Europe (c,d) shows population structure especially within Europe. Based on these results, the imputation status and PCs 1-3 for Europeans and PCs 1-2 for Yoruba were included as covariates in the QTL analyses.

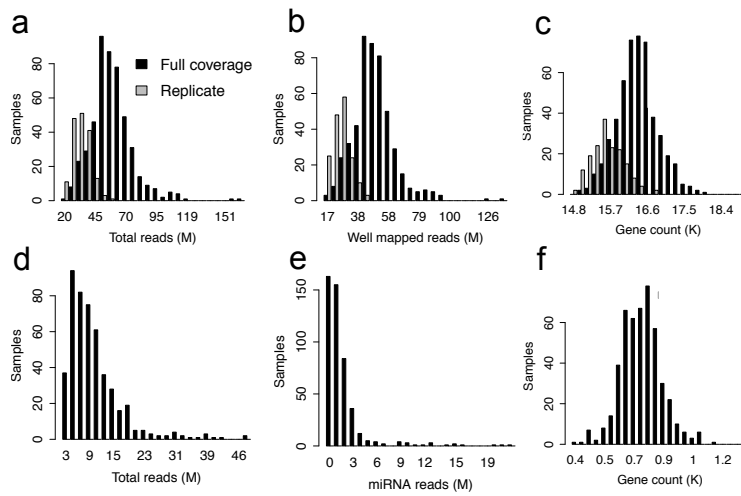


Figure S4. Read and gene count distributions

mRNA statistics per sample of total read counts (a) mapped read counts (MAPQ>150, properly paired, NM<=6) (b), and gene counts (>1 RPKM) (c), and small RNA statistics of total read counts (d) miRNA read counts (e), and miRNA gene (>0) counts (f). These distributions have few outliers, especially in gene counts, demonstrating the uniformity of the raw data. The mRNA replicate samples refer to 168 low-coverage replicate samples that were not used in the analysis.

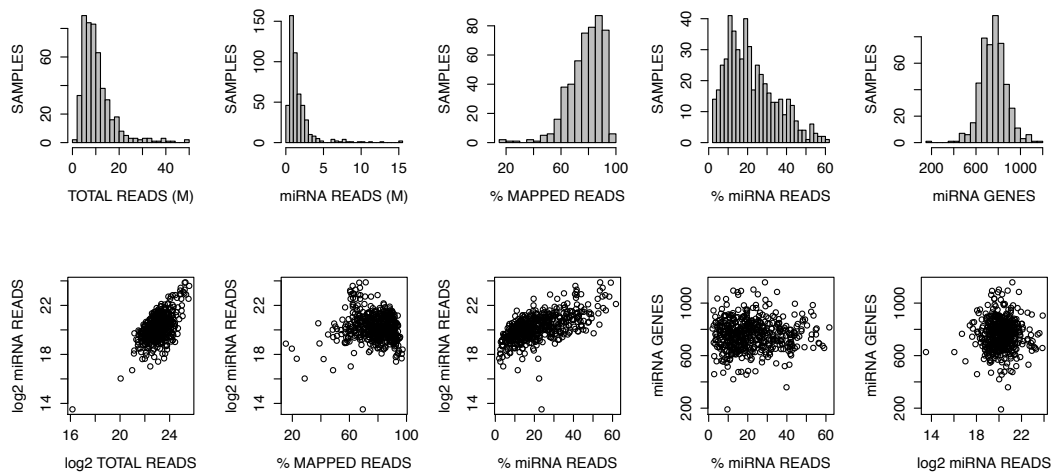


Figure S5. miRNA quality control

Combination of various quality control statistics of small RNA sequencing (total read count, proportion of mapped reads) and miRNA quantification (miRNA read count, proportion of miRNA reads, and number of quantified miRNAs). These plots demonstrate that except for 13 outliers that were excluded from the final data set, the final quantification data is very uniform.

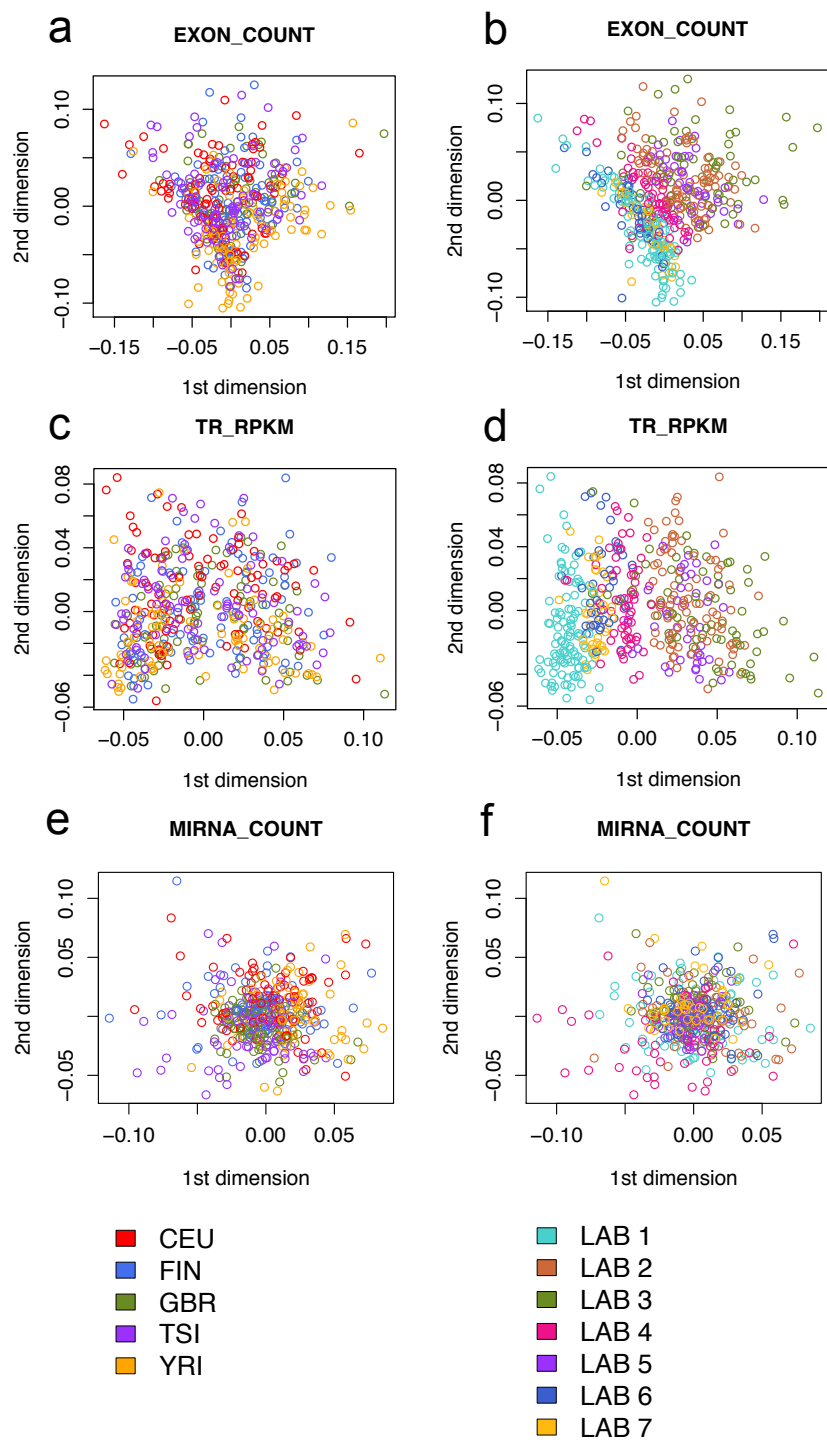


Figure S6. Sample clustering

Multidimensional scaling of pairwise sample correlations based on exon (a, b), transcript (c,d) and miRNA (e,f) quantifications normalized only for the total number of mapped reads. The same data is shown colored by population (a, c, e) and by sequencing laboratory (b, d, f).

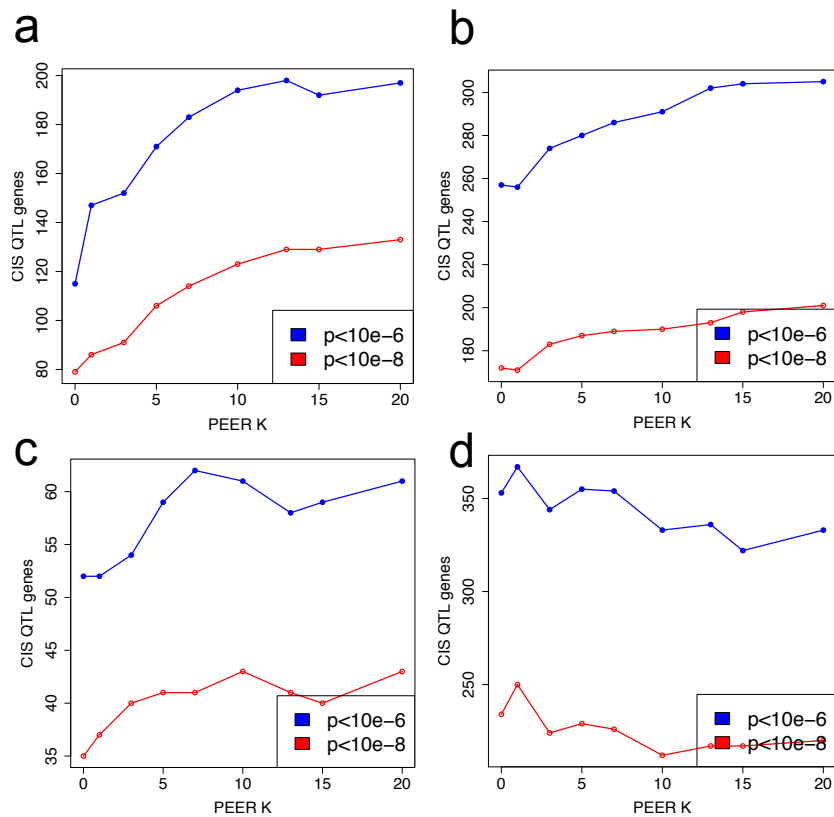


Figure S7. PEER covariate analysis

The number of cis-eQTLs in a small test data set was used to evaluate the performance of PEER normalization for all quantifications; here it is shown as a function of the number of corrected covariates for mRNA (a), transcript (b), miRNA (c) and repeat (d) quantifications. For the final analysis, the data was normalized with $K=10$ except for repeat quantifications for which PEER normalization was not done.

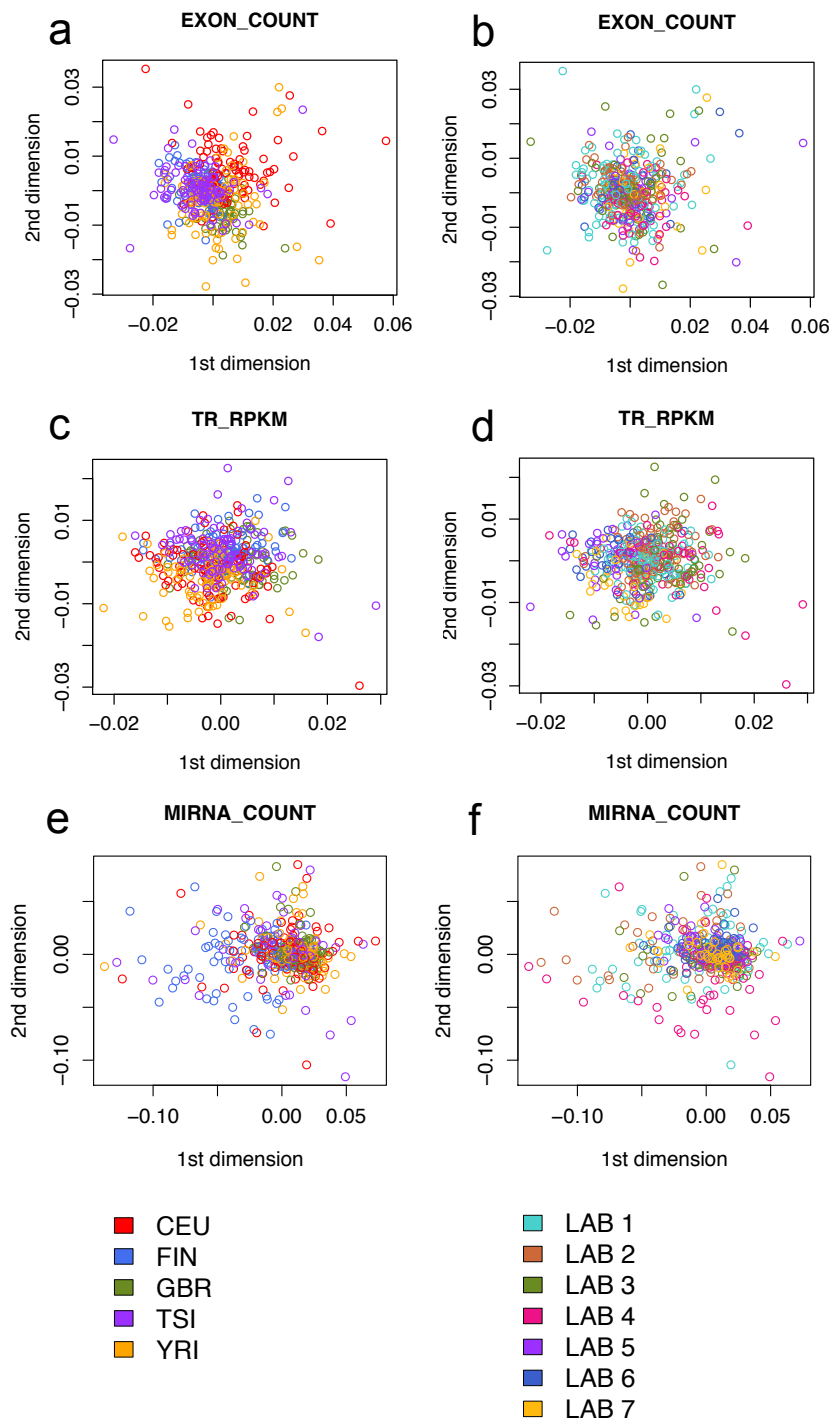


Figure S8. Sample clustering after normalization

Multidimensional scaling of pairwise sample correlations based on exon (a, b), transcript (c,d) and miRNA (e,f) quantifications after PEER normalization. The same data is shown colored by population (a, c, e) and by sequencing laboratory (b, d, f).

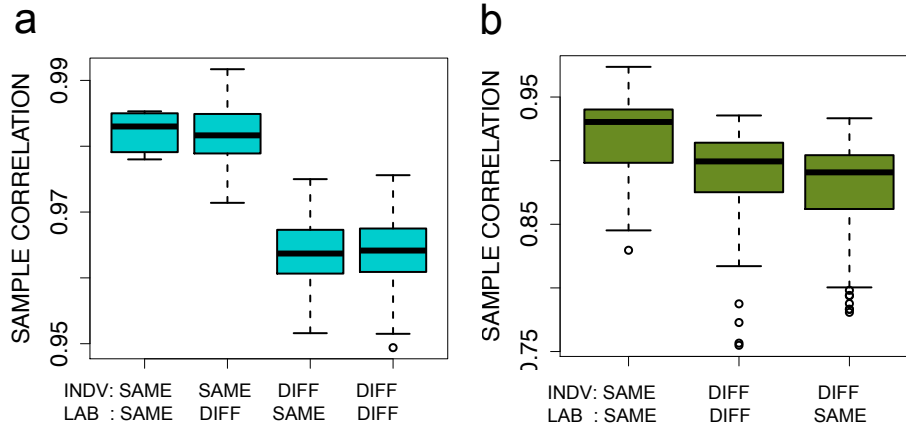


Figure S9. Replicate correlation after normalization

Correlation of the five replicate samples based on mRNA exon (a) and miRNA (b) quantifications after PEER normalization, partitioned by lab and individual. In mRNA sequencing, the same samples were sequenced twice in one lab. See Fig. 1a for similar analysis before normalization.

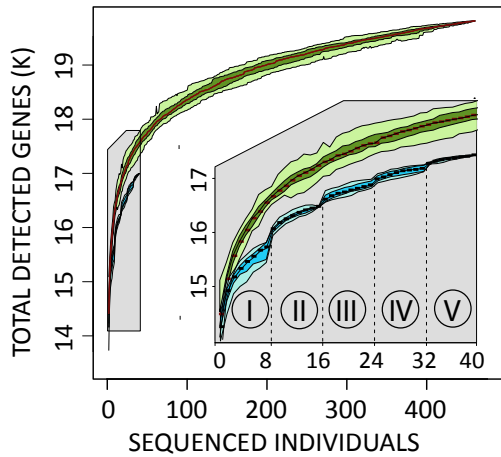


Figure S10. Gene discovery in a population sample

Total number quantified genes (>1 RPKM) across the whole sample set as a function of sequenced samples for nonredundant 462 individuals (green) and for 5 replicate samples sequences 8 times each (blue; roman numbers). The order in which the individuals are added has been permuted 30 times; the thick lines show the medians, and the shaded areas show the 25th and 75th quantiles and the minimum and maximum across permutations. The increasing curve demonstrates how almost every sample expresses some genes not observed in others, and the green curve of different samples being above the replicate samples shows that part of this increase is due to population diversity rather than increased total sequencing depth.

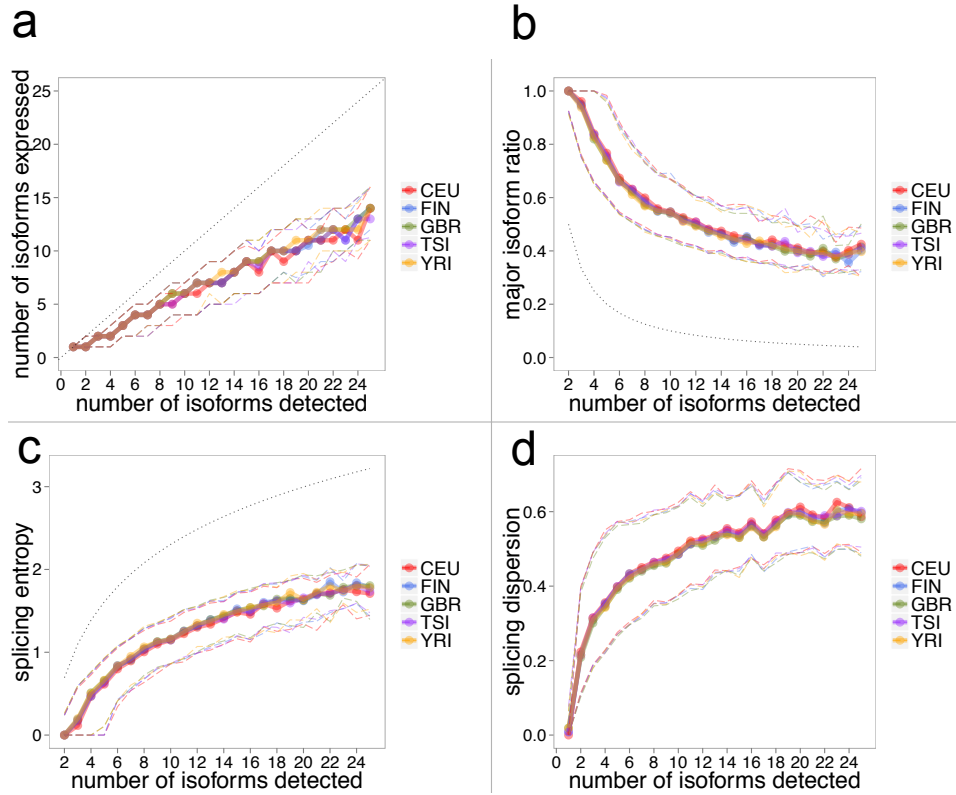


Figure S11. Transcript quantification statistics

Transcript quantification metrics computed at the gene-level, and plotted with medians (points) and first/third quartiles (dashed lines): number of transcripts expressed (>0.1 RMPK) in each population (a), ratio of the major transcript of the total per gene (b), splicing entropy, where one transcript expressed would result in low entropy and all transcripts expressed equally would give a high value (c), and splicing dispersion which represents the variability in the space of the splicing ratios. (d). The genes are grouped according to the number of different transcripts detected when pooling all the samples together (x-axis), as the theoretical boundaries of the metrics depend on the number of transcripts observed (shown as lines in (a) and (c)). The general distribution of these metrics indicates good quality transcript quantifications with expected patterns, as well as high consistency of these basic metrics across the populations.

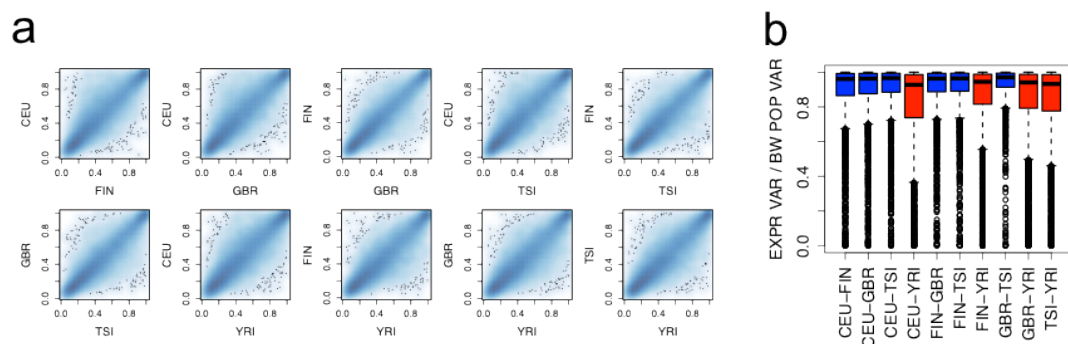


Figure S12. Transcript variation between population pairs

For each gene, we can calculate the proportion of expression level variation (as opposed to splicing) of the total expression variation between individuals in each population, and the comparison of these values is shown in (a). The consistency between populations indicates that each gene has a characteristic pattern of how much expression versus splicing variation it allows. Furthermore, for each gene, a small proportion of total gene expression variation is explained by difference between population pairs. We calculated how much of between-population variation is explained by variation in expression levels, and show this distribution for population pairs for genes with high level of population differentiation (between-population variation >2.5% of total) (b). We observe that population pairs between the African YRI and European populations (in red) have lower proportions of variation explained by expression levels than European pairs (in blue), suggesting a bigger contribution of splicing variation between continental populations. This supports our results from differential expression/transcript usage analysis (Fig. 1c) that indicate higher contribution of splicing differences in between-continent transcript variation.

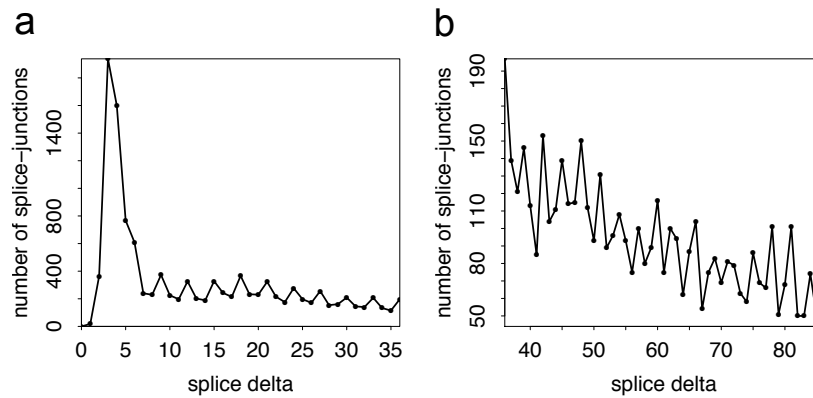


Figure S13. Tandem variation in splicing

Splice-deltas for donor (a) and acceptor (b) sites were calculated as the difference between the location of the major splice-donor or -acceptor and the location of minor splice-donors or -acceptors. Peaks can be observed for $\Delta 3$ splice-sites, which is expected due to common NAGNAG and GYNGYN tandem splice-site patterns, as well as nonsense-mediated decay of transcripts where splice patterns disrupt the reading frame.

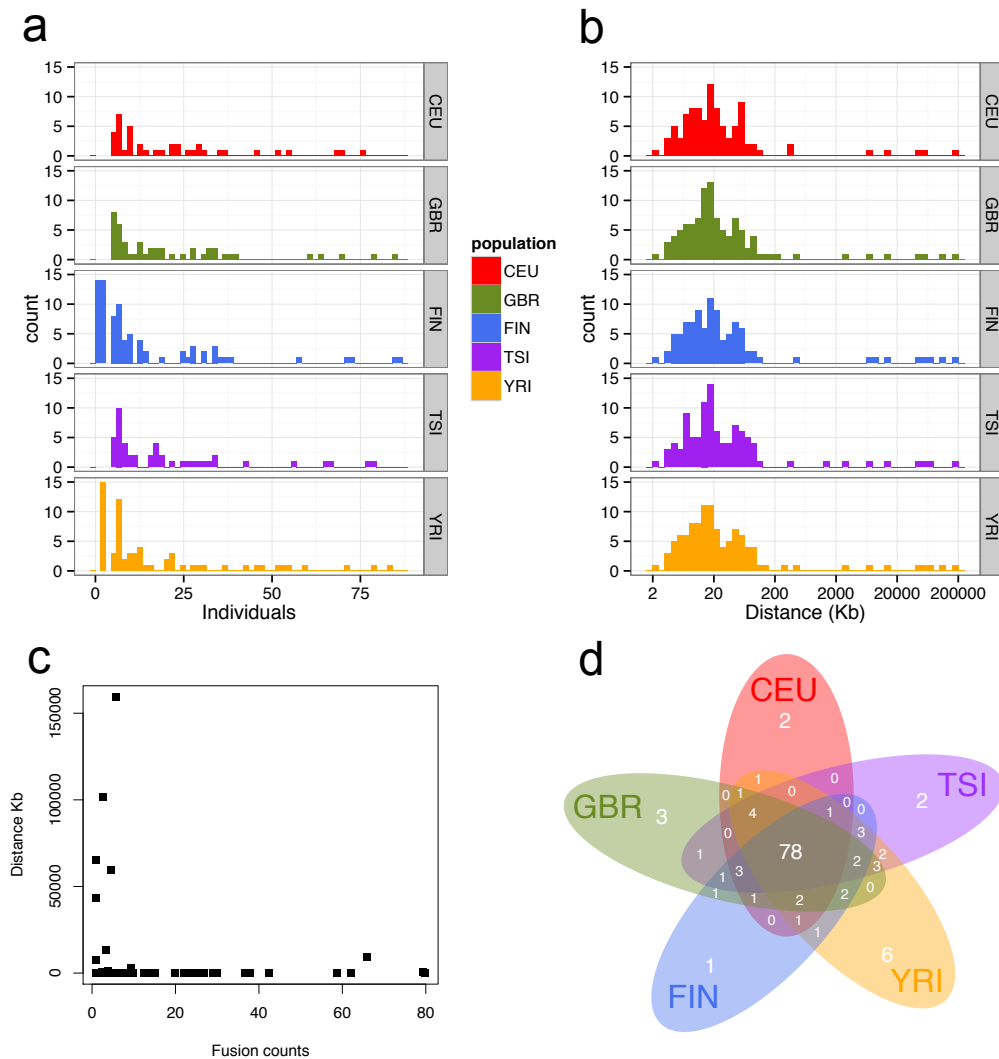


Figure S14. Chimeric transcripts

Frequency distribution of chimeric transcripts that include parts from two annotated genes in the five populations (a) shows that the majority of the chimeric transcripts are present in small number of individuals with only few highly recurrent ones clustered in the right part of figure. Distribution of distance between the breakpoints (b) shows that the range for the distances between the partner genes involved in the fusion varied between 2kb and 200kb except some outliers. Scatterplot of these two (c) confirms that the majority of the chimeric transcripts exhibit small distances between the fusions breakpoint, with a few fusion partners joined at higher distances being mostly rare in the populations. (d) shows the population sharing of detected chimeric transcripts: most of them are observed at least once in all the populations.

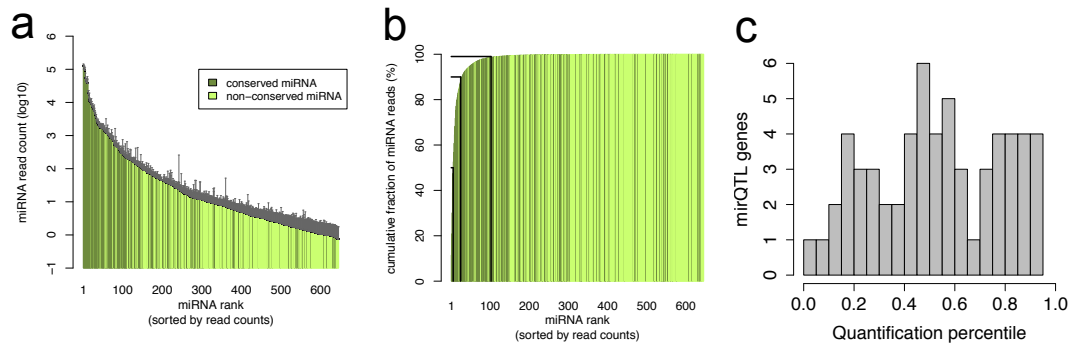


Figure S15. miRNA quantification statistics

Expression of 644 autosomal miRNAs detected in 452 individuals is shown in (a) with mean and s.d. of normalized read counts. (b) shows the cumulative fraction of miRNA reads explained by miRNAs of decreasing abundance. The black lines indicate the 50% fraction (explained by the 6 most abundant miRNAs), 90% fraction (explained by the 29 most abundant miRNAs) and the 99% fraction (explained by the 122 most abundant out of 644 total miRNAs). These plots indicate that the highest expressed miRNAs account for the vast majority of the total miRNA pool of the cell. (c) shows the quantification distribution of miRNAs with mirQTLs, which are found relatively evenly in lowly and highly expressed miRNAs.

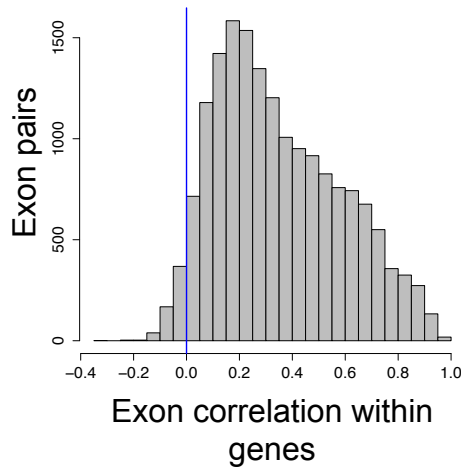


Figure S16. Coexpression of exons of the same gene

Correlation between quantifications of exons from the same gene for chr20 in the European data set (a). For many exon pairs, the correlation is not very high, indicating frequent splicing variation within genes and consistent with the large number of independent eQTL signals for different exons of the same gene.

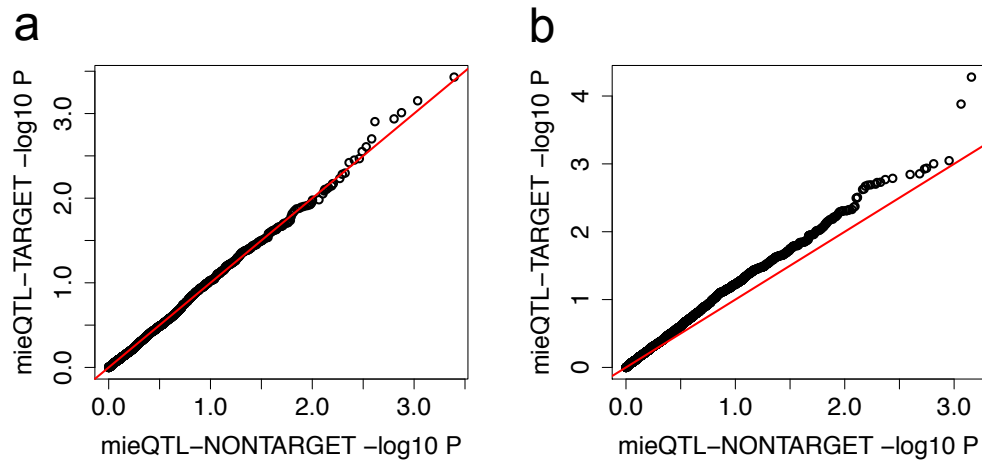


Figure S19. Trans-effects of mirQTLs

Variants that associate to miRNA expression levels can potentially be trans-eQTLs for the target genes of these miRNAs. We sought for this effect by comparing trans-eQTL p-value distributions of cis-mirQTLs to the target exons of the affected miRNA (y-axis) to a null distribution to non-target exons. This comparison was done separately for positive (a) and negative associations (b), i.e. those where the cis-mirQTL allele increasing the miRNA expression has positive or negative correlation to the exon, respectively. The slightly lower p-values for negative associations makes biological sense, since miRNAs are believed to downregulate their targets. While we do not find a long tail of p-values indicating cis-mirQTLs being highly significant trans-eQTLs, the overall shift of the p-value distribution can be a sign of small effects of genetically controlled miRNA levels on their target exons in the population.

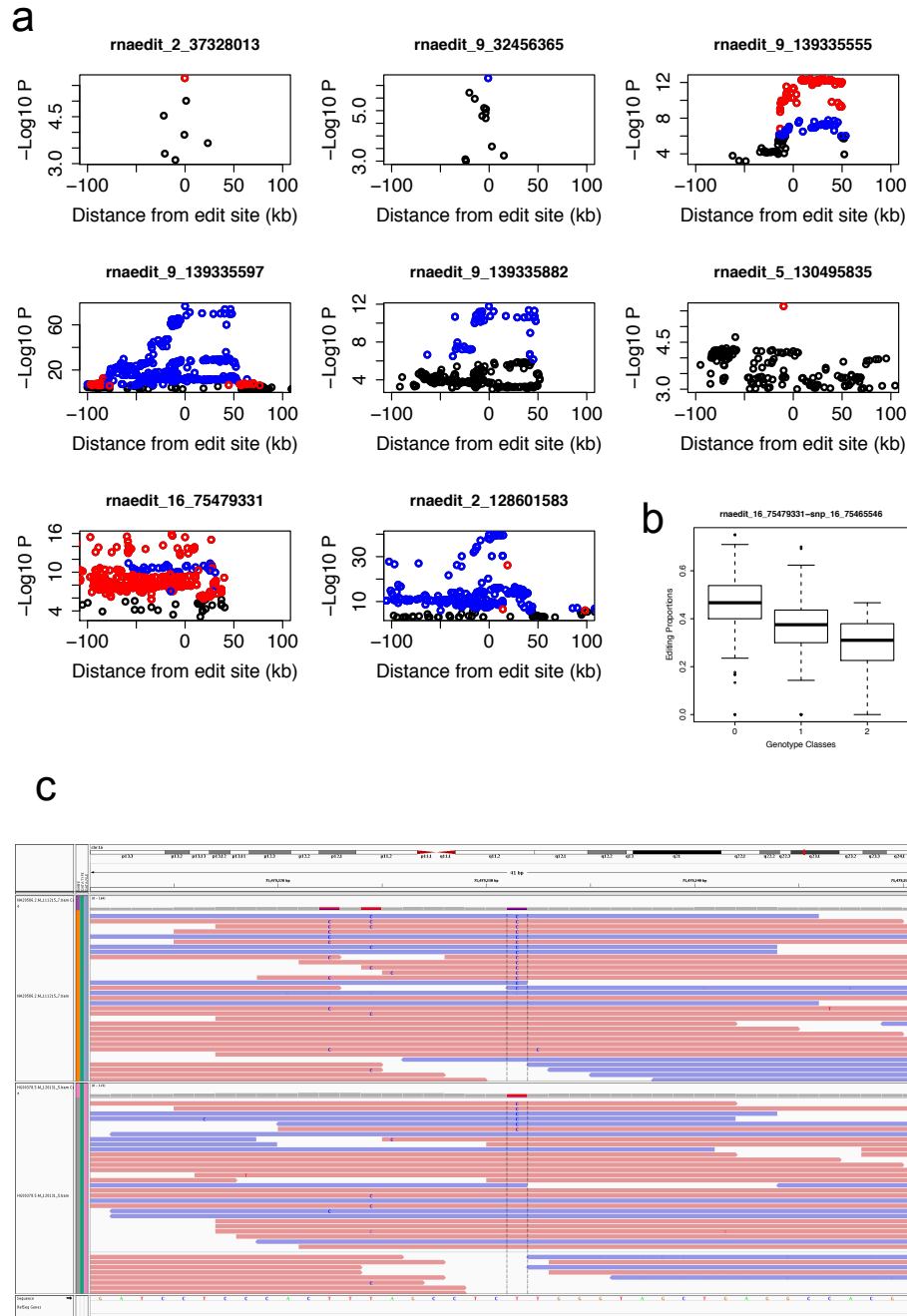


Figure S20. RNA editing QTLs

We detected significant QTLs for the proportion of editing (FDR 5%) for 8 RNA editing sites out of the total 99 quantified sites. (a) shows the genomic landscape of these associations for all variants below editQTL p-value of 0.001, with red and blue showing negative and positive significant associations of the nonreference allele (FDR 5%), respectively. The analysis was done in 1MB window around the editing site, but the figure is zoomed to 100kb window. (b) shows an example of the editing proportion for the genotype classes in rnaedit_16_75479331, and (c) shows the reads overlapping this editing site in two individuals, with the read color showing the orientation and the edited site in the middle, sorted by the allele.

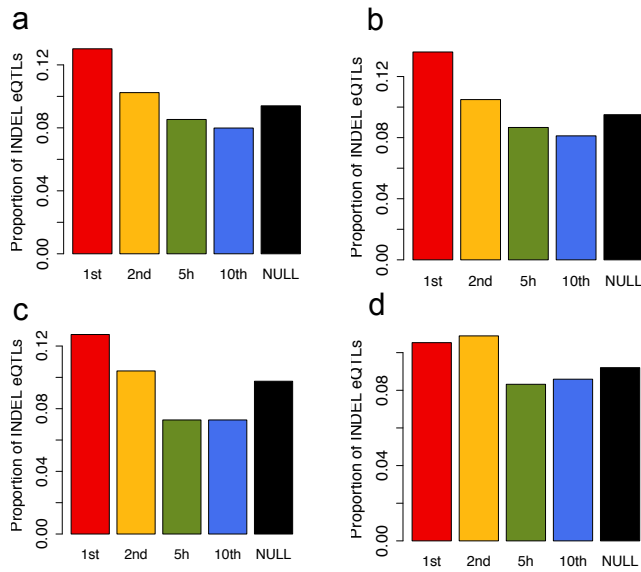


Figure S21. Indel enrichment in eQTL variants

We calculated the proportion of indels among the 1st, 2nd, 5th and 10th best QTL variants and in the matched null for eQTLs in EUR, $p= 5.625e-06$ (a) eQTLs in EUR using only noncoding variants to confirm that the enrichment is not driven by mapping bias, $p= 2.233e-09$ (b) trQTLs in EUR, $p=0.17$ (c), eQTLs in YRI , $p= 0.05796$ (d). We see a clear significant overrepresentation of indels in eQTL and trQTL variants especially in EUR.

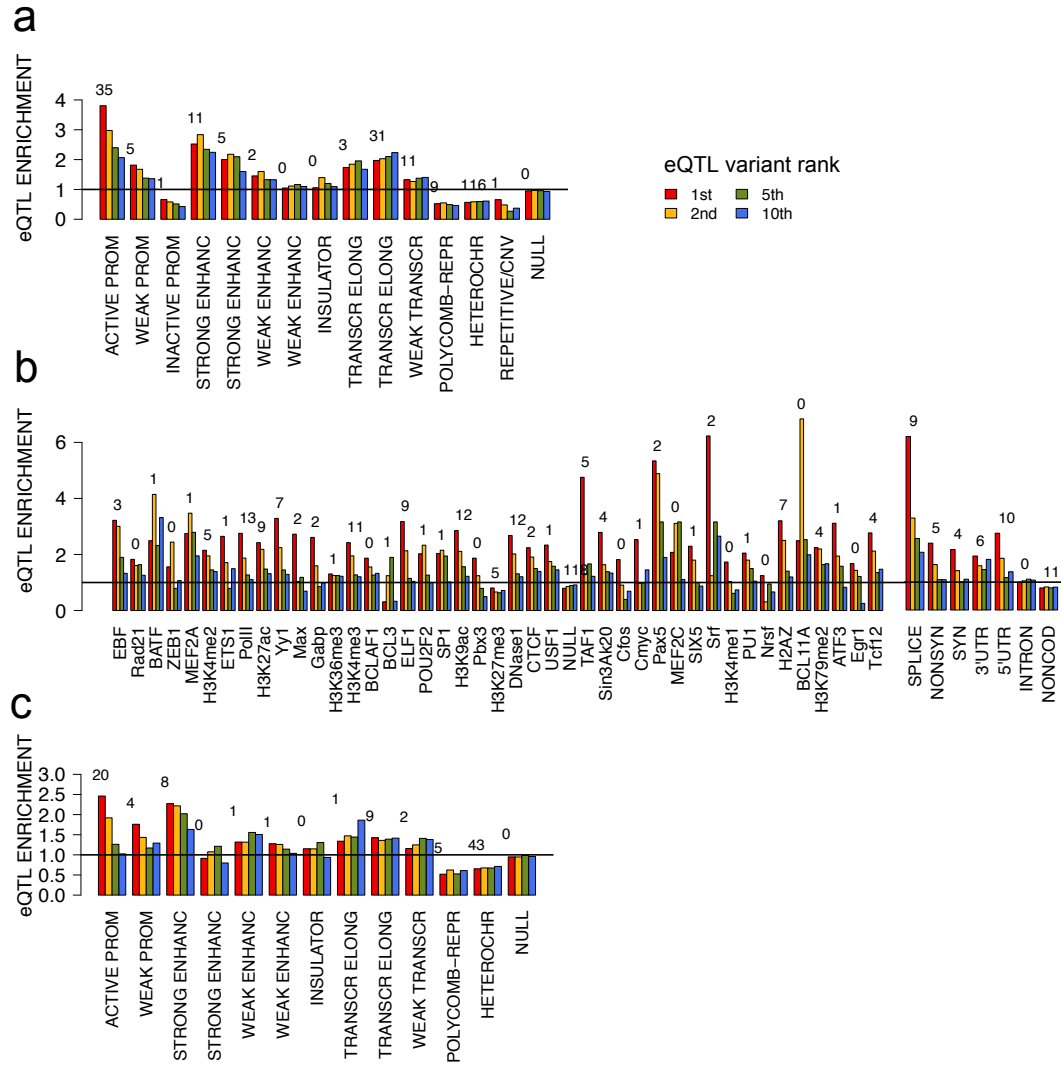


Figure S22. Functional annotation of eQTLs

Enrichment of eQTL variants in functional annotations relative to a matched null distribution for the 1st (most significant) eQTL variant as well as 2nd, 5th and 10th best variants for EUR eQTLs in chromatin states (a), and YRI eQTLs in Ensembl Regulatory Build and coding annotations (b) and in chromatin states (c). See Figure 2a for the figure of EUR eQTLs corresponding to (b) and (c). The numbers above the bars denote $-\log_{10}$ p-values of a Fisher test between 1st eQTL variants and the null for each category. There is an overall high enrichment of eQTLs in functional elements, especially for the 1st variant.

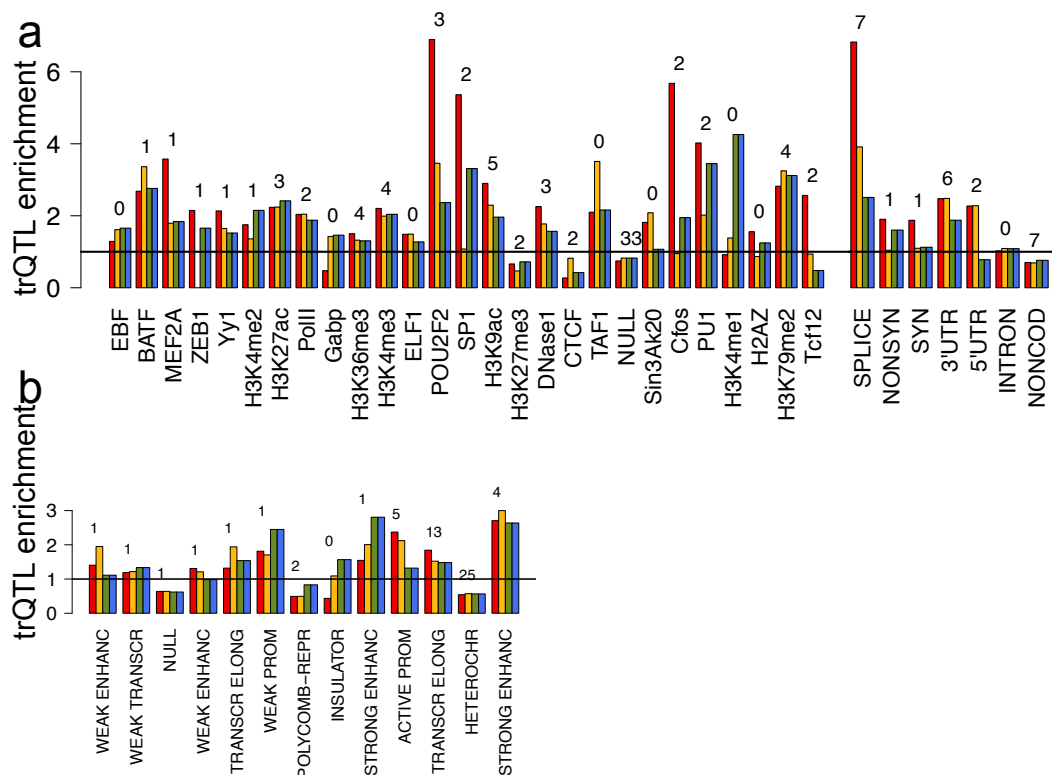


Figure S23. Functional annotation of trQTLs

Enrichment of EUR trQTL variants in functional annotations relative to a matched null distribution for the 1st (most significant) trQTL variant as well as 2nd, 5th and 10th best variants in the Ensembl Regulatory Build and coding annotations (a) and in chromatin states (b). The numbers above the bars denote $-\log_{10}$ p-values of a Fisher test between the 1st the best eQTL and the null for each category. Several annotations are significantly enriched for trQTLs. This analysis is not shown for YRI due to the low number of variants.

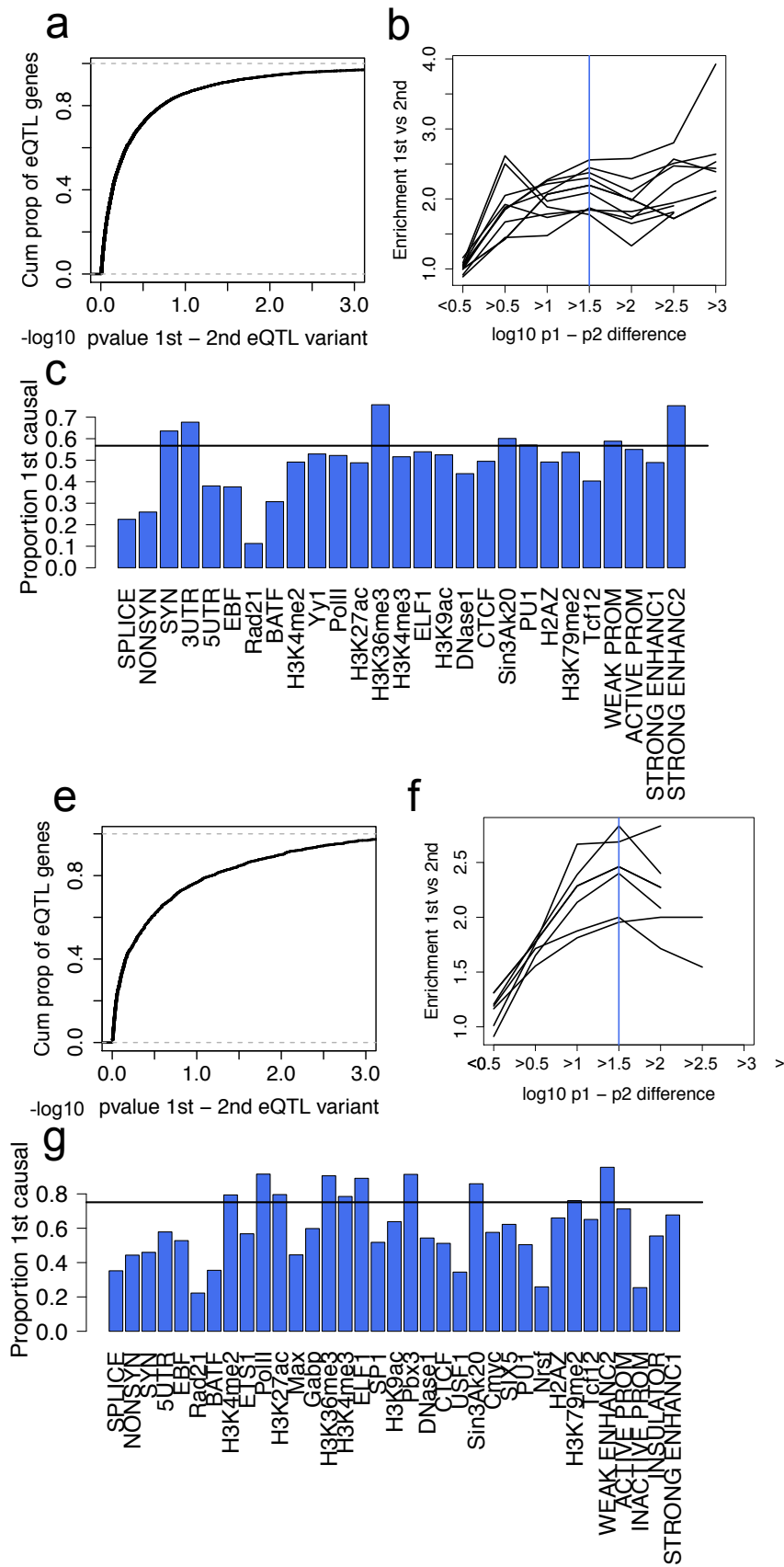


Figure S24. Causal eQTL variants

(From the previous page:) We estimated the probability of the best eQTL variant being the causal regulatory variant by comparing the annotation enrichment (relative to the null) of the best eQTLs variants of all loci to that on loci where the p-value distribution indicates that the best eQTL variant is very likely to be causal. (a-c) show analysis of EUR eQTLs, and the corresponding figures for YRI eQTLs are in panels (d-e). In the majority of eQTLs the p-value difference between the 1st and the 2nd variant (Δp) is small (a,d) due to strong LD between the variants, however, there are also large numbers of eQTLs where the 1st variant association is orders of magnitude more significant than for the 2nd variant, and in such cases the first variant is very likely to be the causal one. We calculated the annotation enrichment of the 1st variant relative to the 2nd variant for different classes of Δp (b,e), and based on the plateau starting at $1-1.5 \log_{10} \Delta p$, we chose $\log_{10} \Delta p > 1.5$ as the limit above which we can safely assume that the first variant is causal. Then, for these causal variants, we calculated the enrichment of the best variants relative to the null, and comparing the same enrichment of all the variants to this number (c) gives us an estimate of the proportion of all variants where the 1st variant is causal, with the weighted median (horizontal line) based on the proportion and frequency of each annotation class.

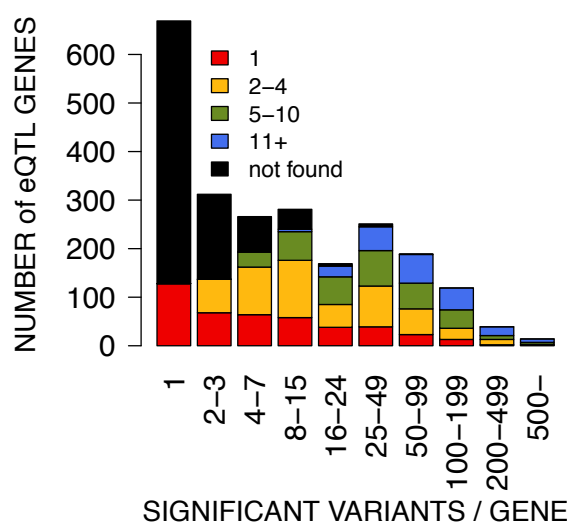


Figure S25. Overlap of eQTLs with Omni 2.5M SNPs

The rank of the best Omni2.5M SNP among the significant YRI eQTL variants per gene, in bins on the x-axis according to the total number of significant variants. See Figure 2 for the plot of European eQTLs.

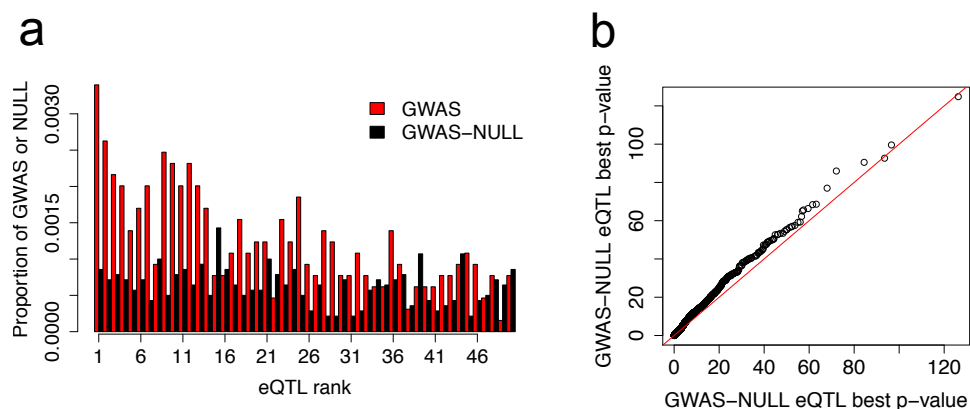


Figure S26. GWAS signal of eQTLs

We analyzed the overlap of GWAS SNPs and EUR eQTL signals. First, we calculated for each GWAS variant its rank among eQTLs (a) – i.e. if the SNP is the best associating eQTL of a gene, it gets a value of one – and this repeated for a null distribution of variants matched to the GWAS minor allele frequencies. GWAS variants are clearly enriched around the peak of eQTL associations compared to the null, suggesting that eQTL variants are the causal variants for many GWAS associations. The distribution is truncated at the rank of 50. (b) shows the best cis-eQTL p-value of GWAS SNPs and the matched null variants plotted as a qq-plot, indicating that GWAS variants are more likely to be eQTLs.

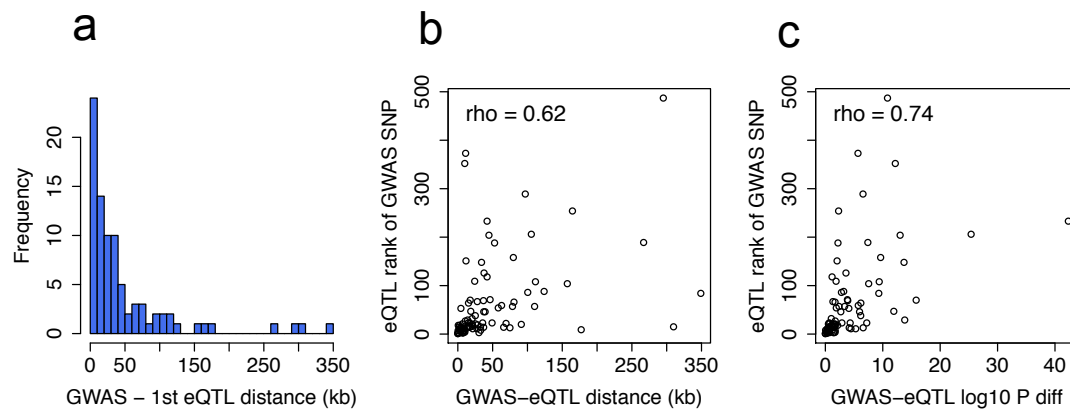


Figure S27. Causal GWAS variants prediction

For the 86 GWAS eQTLs that have been assessed to share a causal variant (see Supplementary Methods), we assessed how close to our best EUR eQTL – the most likely causal variant – the GWAS variant is, in terms of distance (a), eQTL rank versus distance (b), and eQTL rank versus p value difference between the GWAS SNP and the best eQTL (c). 72% of GWAS variants are >10kb away from the most likely causal variant or region, and in the 1000 Genomes data there is a median of 22 variants with better eQTL p-values than the GWAS variant. The correlations in (b) and (c) are significant with $p = 2.279 \times 10^{-10}$ and $p = 4.049 \times 10^{-16}$, respectively.

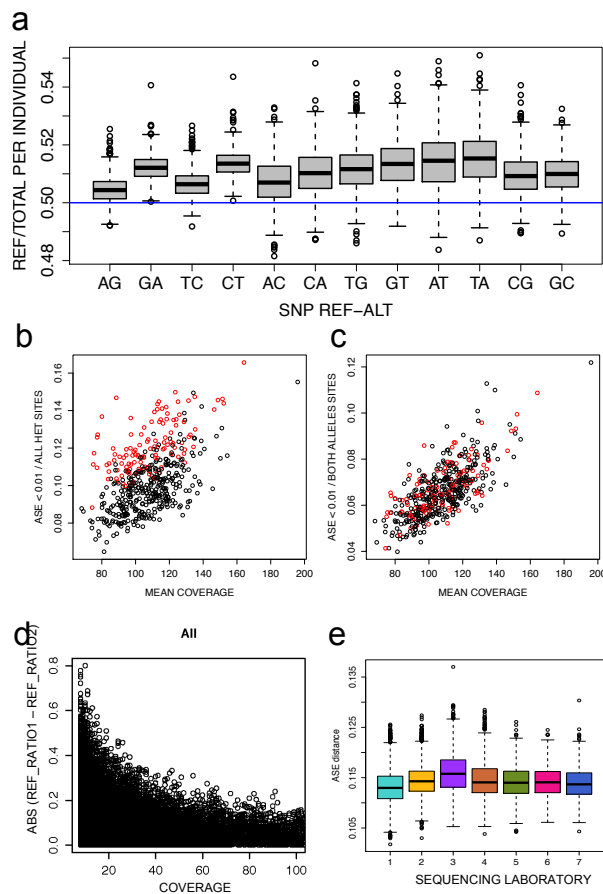


Figure S28. Quality control of ASE data

The expected allelic ratio in ASE analysis (a) is not strictly 50-50 as might be expected for heterozygous sites – there is a slight genome-wide bias favoring the reference allele, and there is also a nucleotide bias favoring G and C, shown in (a) where the genome-wide allelic ratios for each individual are plotted. These ratios are used as the expected ratio in the calculation of binomial probability of ASE.

(b) and (c) demonstrate the effects of slight variation in genotype quality in ASE data. Here, each dot is an individual, with median coverage of ASE sites plotted on the x-axis, and proportion of ASE ($p < 0.01$) on the y-axis. The general correlation between these two is expected, due to higher power when coverage is high. The red individuals are the lowest 25% of DNA-RNA genotype concordance of heterozygous sites – some of which may be due to allelic expression, and some due to false genotype calls. Using these sites in analysis leads to higher proportion of ASE in the red individuals (b), whereas using only sites where both alleles are seen as we did in the majority of analyses (c) corrects for this.

The effect of coverage of the ASE site is demonstrated in (d) using replicate samples: difference in the allelic ratio of the same site is plotted as a function of coverage. Consistency is good after 30-40 reads.

While ASE is less sensitive to laboratory effects than quantifications, ASE distance (see Supplementary Methods) between individuals still shows some laboratory effects.

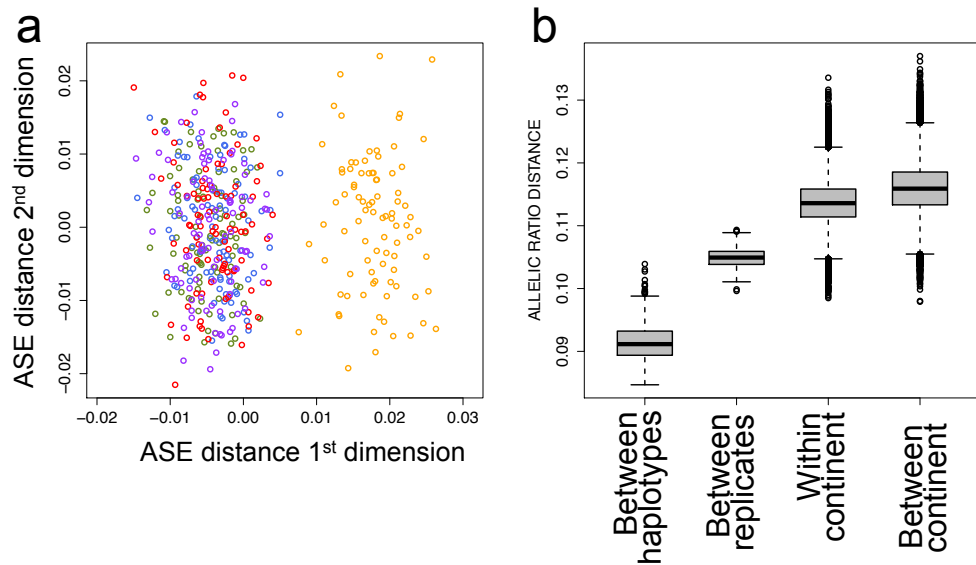


Figure S29. Population variation in ASE

Multidimensional scaling of a matrix of allelic ratio distance between individual pairs (see Supplementary Methods) shows a clear clustering to African and European individuals (a). In (b), we further dissected allelic ratio distances to differences between two haplotypes of an individual ($\text{abs}(0.5 - \text{REF_RATIO})$), and allelic ratio distances for replicate samples from the same individual, two individuals from the continent, or from a different continent.

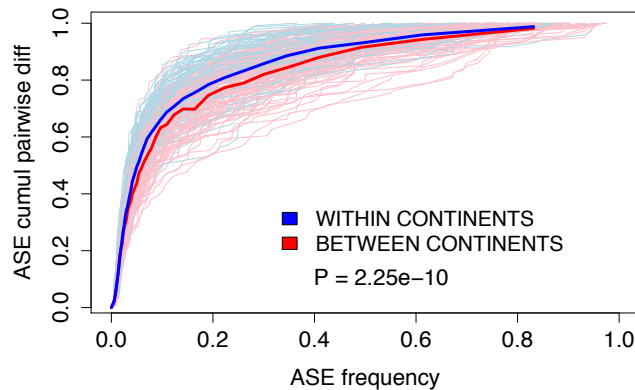


Figure S30. Population variation across ASE frequency spectrum

We used ASE data to partition how much of phenotypic discordance between two individuals come from rare and common events in the population, based on the idea that ASE is a proxy for regulatory variation.

To this end, for each individual pair we took ASE SNPs where the individuals are discordant (ASE $p < 0.005$ & ASE $p > 0.1$). Here, we used only sites present in ≥ 15 individuals in the data set sampled to a coverage of 30, and individual pairs with ≥ 70 sites that were measured in both. For all these sites per individual pair, we calculated the sum of differences in allelic ratios as a measure of total phenotypic difference. Additionally, for each site we calculated how frequent the ASE effect (present in only one of the individuals of the pair) is in the entire sample, which is a proxy for the frequency of the regulatory variant driving this effect.

The plot shows the relationship between the two: the population frequency of ASE (\sim frequency of the regulatory variant) on the x-axis, and the cumulative proportion of how much of the total allelic ratio (\sim phenotypic effect) difference between two individuals explained by each event. Each pair of individuals is represented by a thin line (randomly selected 100 pairs of each class), colors are according to whether the individuals of the pair come from same or different continents. The thick lines represent medians. The p-value is from a Mann-Whitney test of ASE frequencies in within-continent vs between-continent pairs.

We can see that most differences between two individuals are caused by regulatory effects that are rare in the population, which is consistent with the frequency spectrum in Fig. 4. Importantly, differences between individuals of the same continent are relatively more often caused by rare effects. This is consistent with what we know of population sharing of genetic variants: rare variants are very population specific; thus a relatively larger proportion of differences within populations are expected to be driven by rare variants, whereas different continents differ in terms both rare and common variation and so they can contribute more equally to individual differences.

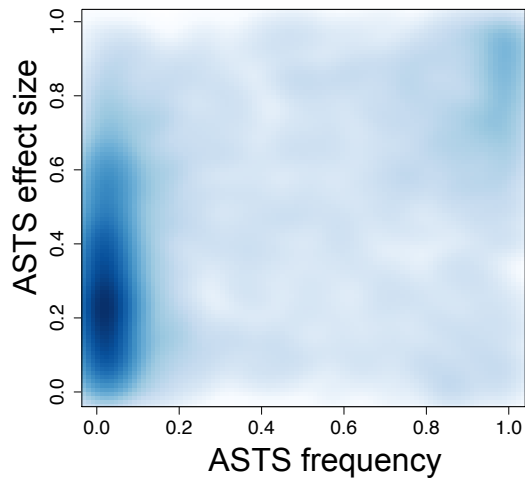


Figure S31. Frequency spectrum of allele-specific transcript structure

ASTS effect size (maximum allelic ratio distance of exons from the total ratio) as a function of frequency of the ASTS effect in the population, calculated for sites with ≥ 20 ASTS measurements. This is analogous to Fig. 3b of ASE, and shows that the majority of ASTS effects are rare in the population.

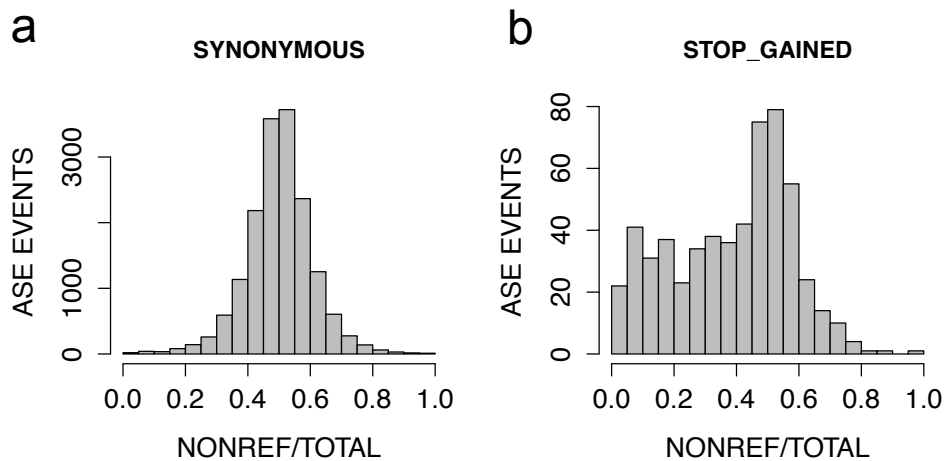


Figure S32. Nonsense-mediated decay

Alternative allele ratio for a random set of synonymous variants in (a) and for stop-gained variants (b), from one random individual per site.

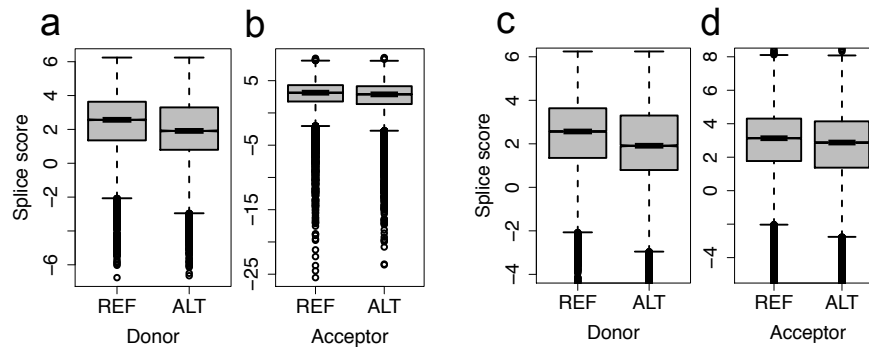


Figure S33. Splice scores

Splicing score predictions of variants overlapping an annotated splice motif. Distribution of scores for reference and alternative alleles for donor and acceptor sites is illustrated, with the plots showing the full distribution for donor (a) and acceptor (b) and the same distributions with zoomed y-axis in (c) and (d). The difference between reference and alternative allele distributions are significant for both donor and acceptor sites ($p < 2.2e-16$).

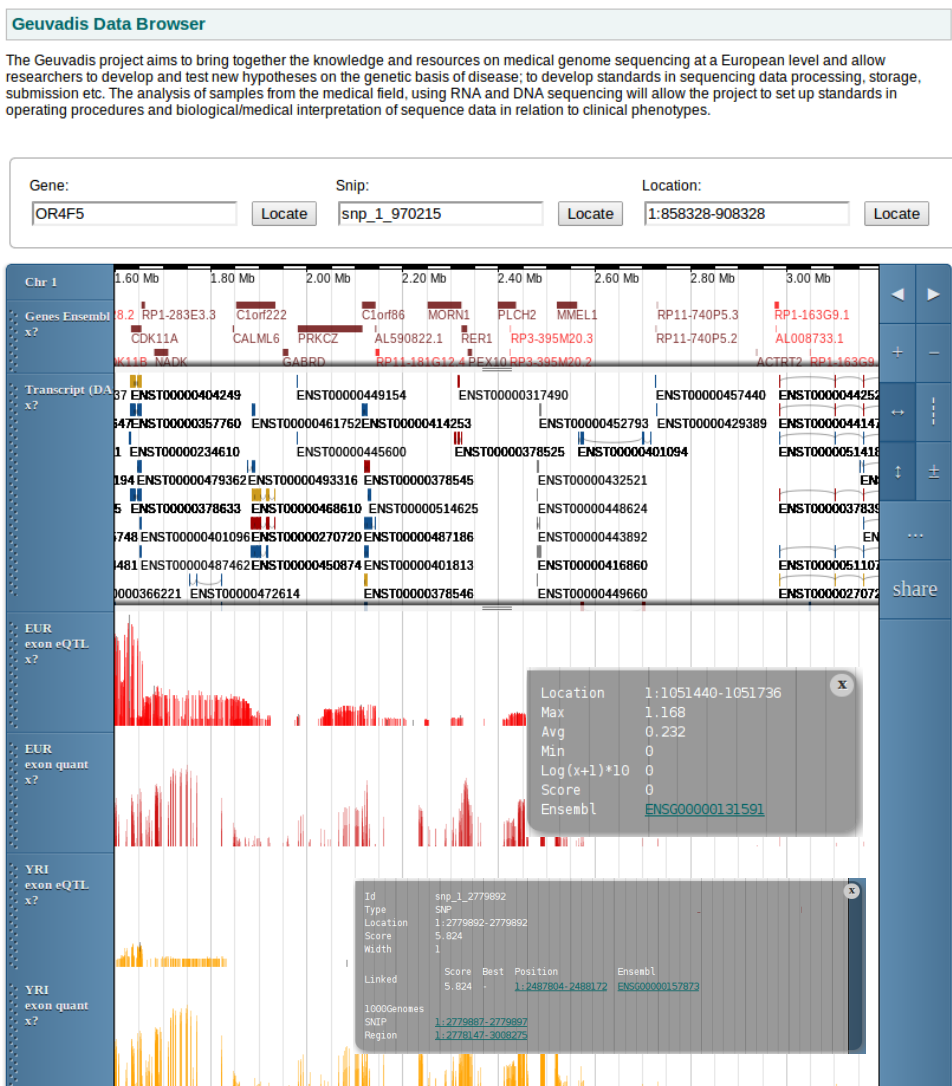


Figure S34. The Geuvadis Data Browser

For the visualisation of RNA-sequencing analysis we created the Geuvadis Data Browser (www.ebi.ac.uk/Tools/geuvadis-das) for viewing exon and miRNA quantifications as well as exon eQTLs and mirQTLs for EUR and YRI. An example view is shown here.

Supplementary Methods

Study design

Transcriptome sequencing was performed in seven European laboratories, each processing 48-116 randomly assigned samples. Five samples were sequenced in replicate in each of the labs for both mRNA and miRNA, and twice at the University of Geneva (UNIGE) for mRNA. Additionally, 168 samples, also sequenced in other laboratories, were mRNA-sequenced at the University of Geneva, at 2/3 of the standard coverage. Of the replicate samples, the one with the highest coverage was used in the main analysis of unique samples.

RNA-sequencing data production

Cell line processing

EVb transformed lymphoblastoid cell lines (LCLs) directly from Coriell Cell Repositories (GBR, FIN, TSI) and originally from Coriell but grown at the University of Geneva (CEU, YRI) were shipped to ECACC (European Collection of Cell Cultures) as live cultures, in batches of ~30 samples from Coriell (GBR/FIN/TSI somewhat randomized) and 2 x ~90 samples (by population) from Geneva.

In ECACC, these cell lines were cultured to approximately 1.2×10^8 cells. These cultures were split to produce 8 x cell banks of the samples, and a snap frozen pellet of 2×10^7 cells from a proliferating culture. The cell pellets were shipped from ECACC to University of Geneva in three batches, the first batch consisting of CEU/GBR/FIN/TSI samples, and the second and third batch with YRI and the rest of CEU samples.

RNA extraction

RNA was extracted in Geneva about 14 samples at a time, first extracting 2/3 of the first shipping batch with full randomization, then adding the second batch and randomizing among that and the remaining 1/3 of the first batch, and finally extracting the third batch.

Total RNA was extracted from cell pellets using the TRIzol Reagent (Ambion). The pellets had been frozen at ECACC without any additives like RNAlater or TRIzol. In Geneva they were thawed, 1mL of TRIzol was added in each sample, and the samples were transferred to eppendorf tubes. The rest of the protocol followed the manufacturer's guidelines. No DNase treatment was done to the RNA samples.

RNA quality was assessed by Agilent Bioanalyzer RNA 6000 Nano Kit according to the manufacturer's instructions. RNA quantity was measured by

Qubit 2.0 (Invitrogen) using the RNA Broad range kit according to the manufacturer's instructions.

RNA sequencing

Each of the sequencing laboratories were sent a minimum of 4 ug of total RNA of the samples allocated to them, and RNA Bioanalyzer was ran for 10-20% of the RNA samples before library preparation to confirm sample quality after shipping. No further purification steps were done to the RNA samples other than that specified in the sequencing protocols. Library preps were done in random order in every laboratory.

mRNA sequencing was done on the Illumina HiSeq2000 platform with 75 bp paired-end sequencing with fragment size of ~280 bp – some laboratories sequenced 100bp reads, which were trimmed to 75bp. TruSeq RNA Sample Prep Kit v2 (the high-throughput protocol) was used for library preparation, TruSeq PE Cluster Kit v3 for cluster generation, and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired minimum of 10M mapped and properly paired read pairs from any standard mapper, without filtering for mapping quality.

Small RNA sequencing was done on the Illumina HiSeq2000 platform with 36 bp single-end sequencing with fragment size of 145-160 bp. Some laboratories sequenced 50bp reads which were trimmed to 36bp. TruSeq Sm RNA Sample Prep kit was used for library preparation, TruSeq PE Cluster Kit v3 for cluster generation, and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired minimum of 3M total reads.

Extensive information of sample processing was collected from all the laboratories for both mRNA and miRNAseq in order to enable control of batch effects.

Raw data processing

Each lab submitted one demultiplexed fastq file per sample per mRNA and miRNAseq, produced by CASAVA 1.8 or 1.8.2 allowing one mismatch in the index. Reads failing Illumina quality filtering were removed. The fastq files are named as: SAMPLE_ID.SeqLabNumber.M/MI_YYMMDD_Lane_Read.fastq.gz, where M/MI stands for mRNA or miRNA sequencing, and YYMMDD is the sequencing date. All the data were submitted and initially stored in the project ftp site. Samtools²⁸ was used for general data processing throughout the project.

Genotype data

Since not all 1000 Genomes variants have rs-identifiers, we renamed all the variants as follows: SNPs had an identifier of type snp_chr_pos (e.g. snp_21_357682), and indels and structural variants were of type indel/sv:lengthI/D_chr_startpos (indel:3D_1_10523). A key for the conversion of the variants that have an rsID are provided with the genotype and eQTL data files.

Variant annotation

The Variant Effect Predictor (VEP v2.5; http://useast.ensembl.org/info/docs/variation/vep/vep_script.html) tool from Ensembl was modified to produce custom annotation tags and additional loss of function (LoF) annotations. The additional LoF annotation was applied to variants that were annotated as STOP_GAINED, SPLICE_DONOR_VARIANT, SPLICE_ACCEPTOR_VARIANT, and FRAME_SHIFT and flagged if any filters failed. A LoF variant is predicted as high confidence (HC) if there is at least one transcript that passes all filters, otherwise it is predicted as low confidence (LC). This modified version of VEP was applied to the 1000 Genomes Phase1 data using the Gencode v12 annotation. To this, we added information of overlap with chromatin states²⁹, Ensembl Regulatory Build elements, miRNA targets from TargetScan³⁰, and miRBase v18³¹ mature and hairpin miRNA loci. Annotation information is stored in the vcf file info field as ordered lists. Detailed documentation is provided together with the vcf files.

Imputation

For 421 samples of the project, we used the 1000 Genomes Phase1 release v3. Genotype imputation was done for 42 samples from 1000 Genomes project Phase 2 with Omni 2.5M genotype data, using the IMPUTE2 software³². As the reference panel we used the entire Phase 1 v3 release, and for a study panel we took Omni Shapeit haplotypes for the whole Phase 2 sample set, and extracted our 42 samples from the imputation results. These were merged to a single vcf file together with the Phase 1 samples.

Since IMPUTE2 did not handle multiallelic genotypes well, we kept only biallelic genotypes for the analysis. Additionally, the genotype calls of imputed genotypes with posterior probability <0.9 were marked as missing.

Quality control

First, we calculated an IBS matrix of genotype data of chr20, which showed clear clustering to Phase1 and Phase2 individuals (Fig. S3), even though we verified that all variants had consistent allele frequencies. Furthermore, PCA³³ showed a clear clustering to populations, as expected. To make sure that our findings are not driven by biases from imputation or from population structure, we included the imputation status (0|1) and principal components 1-3 for Europeans and 1-2 for Yoruba as covariates in QTL analyses.

In QTL analyses, we used variants with >5% MAF in either EUR or YRI, which gave us 10,785,347 variants in total, of which 9,836,718 are SNPs, 945,987 are indels, and 2642 are SVs. QTL analysis was done with genotype dosage values.

mRNA read mapping

We employed the JIP pipeline (Griebel & Sammeth submitted) to map RNA-Seq reads and to quantify mRNA transcripts. For alignment to the human reference genome sequence (GRCh37, autosomes + X + Y + M), we used the GEM mapping

suite (v1.349 which corresponds to publicly available pre-release 2)³⁴ to first map (max. mismatches = 4%, max. edit distance = 20%, min. decoded strata = 2 and strata after best = 1) and subsequently to split-map (max.mismatches = 4%, Gencode v12 and *de novo* junctions) all reads that did not map entirely. Both mapping steps are repeated for reads trimmed 20 nucleotides from their 3'-end, and then for reads trimmed 5 nucleotides from their 5'-end additionally to earlier 3'-trimming—each time considering exclusively reads that have not been mapped in earlier iterations. Finally, all read mappings were assessed with respect to the mate pair information: valid mapping pairs are formed up to a max. insert size of 100,000 bp, extension trigger = 0.999 and minimum decoded strata = 1. The mapping pipeline and settings is described below, and can also be found in <http://github.com/gemtools>, where the code as well as an example pipeline are hosted.

The GEM output format was converted to bam format, with following mapping quality scores and flags:

1. Matches which are unique, and do not have any subdominant match: 251 \geq MAPQ \geq 255, XT=U
2. Matches which are unique, and have subdominant matches but a different score: 175 \geq MAPQ \geq 181, XT=U
3. Matches which are putatively unique (not unique, but distinguishable by score): 119 \geq MAPQ \geq 127, XT=U
4. Matches which are a perfect tie: 78 \geq MAPQ \geq 90, XT=R.

Furthermore, the NM flag contains the number of total mismatches (read1+read2). In analysis, we used reads in categories 1 and 2 and with NM \leq 6.

The analysis of chimeric transcripts was based on read mapping with bwa-0.5.9³⁵ with default parameters.

mRNA quantifications

The gene annotation used in this project was Gencode v12³⁶.

Exons and genes

Exon quantifications were calculated for protein-coding and linc-RNA transcripts. All overlapping exons of a gene were merged into meta-exons with identifier of type ENSG000001.1_exon.start.pos_exon.end.pos. Read counts over these elements were calculated without using information of read pairing, except for excluding reads where the pairs map to two different genes. We counted a read in an exon if either its start or end coordinate overlapped an exon. For split reads, we counted the exon overlap of each split fragment, and added counts per read as 1/(number of overlapping exons per gene). Gene counts were calculated as the total number of reads overlapping exons of each gene.

Transcripts, splice junctions, and introns

Quantifications of transcripts, introns and splice-junctions by the Flux Capacitor approach³⁷ are based on the annotation-mapped genomic mappings considering

transcript structures of the Gencode transcriptome annotation: mappings of read pairs that were completely included within the annotated exon boundaries and paired in the expected orientation have been taken into account. For intron quantifications, we used all-intronic regions that are not retained in any mature annotated transcript, and reported mapped reads in different bins across the intron in order to distinguish reads stemming from retained introns from those produced by not yet annotated exons. Reads belonging to single transcripts were predicted by deconvolution according to observations of paired reads mapping across all exonic segments of a locus. Annotated splice junctions were quantified using split read information, counting the number of reads supporting a given junction. Novel splice sites were quantified by split-reads that are overlapping a window of +/-30 bp around known exon boundaries. Each different pattern of split-reads was regarded as a potential splice-isoform, if they had insert-size and split-distance of at most 10,000 bp. Gene quantifications were calculated as the sum of all transcript RPKMs per gene. Transcript ratios were calculated as the proportion of each transcript quantification (in RPKM) of the sum of all transcripts per gene.

Exon inclusion

Exon inclusion levels were calculated as the Percentage Splice Index (PSI)^{38,39}, defined as the ratio between inclusion reads and inclusion reads plus exclusion reads.

Transcribed repeats

We quantified transcription on repeat elements using the following approach: First, we extracted all repetitive elements from UCSC's repeat masker table, and excluded all elements that overlapped UCSC or Gencode genes by at least one nucleotide. This left us with 2.5M regions, in which we then counted the number of overlapping RNA-seq reads in each region for each sample. Reads that were partially overlapping are only counted for the part which is overlapping. Since we observed that rRNA elements had strong differences between laboratories, we excluded them from further analysis.

small RNA (sRNA) data processing

Improved miRNA gene annotations

Our annotation builds on miRBase version 18³¹ but with important improvements. In the cases where only one miRNA strand was annotated, the position and sequence of the other strand was estimated using RNA structure prediction⁴⁰. Furthermore, for the mature and hairpin miRNAs which overlapped SNP or indel variants that were polymorphic in our genotype data, sequences carrying the nonreference alleles were generated and used for downstream analyses together with the reference sequences. This is important for avoiding allelic mapping bias that can easily occur for short sequences.

sRNA read data processing

Small RNA reads with homo-polymer and low PHRED scores were removed. Ligation adapters were clipped using the AdRec.jar program from the seqBuster suite⁴¹ with the following options: `java -jar AdRec.jar 1 8 0.3`. A custom search subsequently clipped shorter adapters: if there were no matches to the first 8 nts, then matches to the first 7 nts of the adapter were searched in the last 7 nts of the read, then matches of the first 6 to the last 6 positions and so on. Reads that had no matches were retained, but not clipped. Last, reads shorter than 18 nts were discarded.

sRNA mapping and quantification

For tracing the reads to their genomic source for quality control purposes, reads were mapped to the hg19 genome concatenated with unassembled parts of the human genome and genomes of known human viral pathogens (available upon demand) with this command line: `bowtie -f -v 1 -a --best --strata`. sRNA reads were assigned to annotations based on the genome mappings. Annotations used were from Gencode v8³⁶ supplemented with rRNA and LINE and Alu transposon annotations from RepBase⁴² and snoRNA⁴³ and miRNA³¹ annotations. Annotations were first resolved so that each nucleotide on each strand had exactly one annotation. In case of nucleotides with more than one annotation, conflicts were resolved using a confidence-based floating hierarchy (as in ⁴⁴): mitochondrion > virus > miRNA > snoRNA > rRNA > tRNA > snRNA > misc_RNA > lincRNA > processed_transcript > pseudogenes > protein_coding > LINE > Alu > intron_coding > intron_non_coding > intergenic. Each read mapping was weighted inversely to the number of genome mappings for the read, e.g. a read mapping to two genomic locations would get an assigned weight of 0.5. Each mapping was counted towards the annotation of the nucleotide in the middle of the mapping.

miRNA quantifications for analysis were calculated as read counts using miraligner.jar from the seqBuster suite using the following options: `java -jar miraligner.jar 1 3 1`, and using the improved annotations as the reference. Reads that mapped equally well to two or more miRNAs are counted fully towards each miRNA.

RNA-seq quality control

A more detailed analysis of technical variation of this dataset can be found in 't Hoen et al. (in preparation).

Outlier detection

The read and gene count distribution of mRNA-seq data were very uniform (Fig. S4). To further estimate sample quality, we calculated Spearman rank correlation between all samples using exon counts and transcript RPKMs. From these data, we calculated the median correlation of one sample against all the

other samples. 2 samples in mRNA data and 4 samples in miRNA data were excluded from analysis due to low correlation with other samples.

We also used multidimensional scaling to visualize the sample correlation matrix. The samples clustered relatively uniformly but with some separation by the sequencing lab, but this effect was completely removed by normalization (see below; (Fig. S6,S8,S9).

Sample swap and contamination analysis

Allele-specific expression (ASE) analysis of mRNA-seq data was used to detect sample swaps, which we did not find. Based on analysis of increased heterozygosity in ASE results and mixed expression pattern of sex chromosome specific genes, we excluded 5 samples because of possible contamination (t Hoen et al. in preparation).

miRNA data quality control

The total small RNA read count and the number of miRNA reads were relatively similar across samples, but the proportion of miRNA reads per sample showed large variation from close to 0 to 60% (Fig S4,S5). This is likely caused by variation in the library preparation step and sequencing of a large number of non-miRNA reads in some samples. However, the number of quantified miRNAs is very uniform, and is not correlated to the proportion of miRNA reads (Fig S5), and only 8 samples were excluded due having low mapping rate, coverage, or gene count. This indicates that while in some samples sequencing depth is lost on non-miRNA reads, this hardly affects our miRNA detection and quantification. Notably, correlations between miRNA samples were high, and population clustering clearly more pronounced than clustering by laboratory even before normalization (Fig 1a, S6, S8, S9).

Normalization of quantifications

All read count quantifications were corrected for variation in sequencing depth between samples by normalizing the reads to the median number of well-mapped reads (45M) for mRNA, and to the median number of miRNA reads (1.2M) for miRNA. In general, we used only elements quantified in >50% of individuals unless mentioned otherwise (Table S3).

All expression quantifications are affected by technical noise that reduces power, and it has been shown in many studies that correcting for such sources of variance improves eQTL discovery dramatically. We normalized quantification data using PEER⁴⁵, which finds synthetic covariates from quantification data that can then be regressed out. These normalizations were done for the total sample set as follows:

First, for each type of quantifications, we estimated the best number of covariates (K) to correct: PEER was ran for a subset of the data (chr20, or chr20-22) using K=0,1,3,5,7,10,13,15,20, and sequencing lab and population as additional covariates, the resulting corrected quantifications were transformed to standard normal distribution, and cis-eQTL analysis was performed for each

K. The number of genes with an eQTL ($p < 10^{-8}$ and $p < 10^{-6}$) was calculated, since eQTL discovery is a good indicator of power to find biological effects. These results can be seen in Figure S7.

Based on these results, we chose $K=10$ as the number of covariates to correct for, except for transcribed repeats where we did not use PEER correction. To normalize the final data sets, we ran PEER for 20 000 quantification units (e.g. exons) using sequencing laboratory and population as additional covariates and adding the mean to the model. Covariates from this analysis were regressed out from all the quantifications, and the mean was added to the residuals. Correlation of samples after this normalization showed no remaining laboratory effects for mRNA data across the samples and in replicates (Fig. S8, S9), and while these effects were not completely removed for miRNA (probably due to smaller number of genes which allows less efficient calculation of synthetic covariates), they are hardly visible in sample clustering. In eQTL analysis and miRNA-mRNA correlation analysis, these quantifications were further transformed to standard normal distribution.

Quantitative versus qualitative variation

We estimated the contribution of alternative splicing and gene expression on the total transcript abundance variation using approaches in Gonzales-Porta et al. 2012⁴⁶. Briefly, for each gene, the samples are represented in the R^T space using transcript expression levels (T =number of expressed transcripts for this gene), from which we can calculate the total variability (V_t) in this space. Projecting the samples in a model of constant splicing ratios gives us an estimate of expression level variation (V_l s). The ratio V_l s/ V_t estimate the contribution of gene expression in the transcript abundance variability, where V_l s/ $V_t \approx 1$ implies that only gene expression contributes to transcript variability, and V_l s/ $V_t \approx 0$ implies that only transcript usage variation contributes to transcription variability of the gene. In this analysis, we used only protein-coding genes expressed in at least 20 individuals per population with at least two expressed (RPKM ≥ 0.01) transcripts.

We further extended this model to between-population variation. Representing the samples in the space of the transcript expression, between-group variation was computed removing the within-group variation from the total variation. Then all the samples were projected on a line, which represents the model of constant splicing ratios. The between-group variation of these projected points was computed, and the estimator of gene expression level variation between populations is the ratio of between-group variation of the projected points over between-group variation of the original points. A value close to one means that the projection did not remove variation, so gene expression is the one mainly contributing to between-population variation.

Differentially transcribed genes

We performed gene differential expression (DE) analysis using *tweedEseq*⁴⁷, a method that uses a Poisson-Tweedie family of distributions and is well suited to

compare groups with more than 15 samples. 16,583 genes with more than 5 counts per million in at least 1 sample were analyzed in pairwise population comparisons. Genes with $FDR < 0.05$ and \log_2 fold change greater than 2 were considered significant.

We applied the Wilcoxon-Mann-Whitney rank sum test to identify transcripts with significantly different relative abundances between population pairs. The p-values of the individual comparisons were adjusted using the FDR method. Genes with differential transcript usage were then obtained by mapping those transcripts to the associated gene ids.

As a next step we integrated the results from the differential gene expression and differential transcript usage analyses to identify genes that differ only at the level of expression (thus pointing to a transcriptional mechanism as the cause for the observed differences), genes that differ only at the level of transcript usage (thus indicating that the observed differences are due to changes in splicing) and genes that differ in both. Only protein coding genes were used in this analysis.

Chimeric transcripts

We discovered chimeric transcripts with parts from two genes using the following pipeline: Unmapped reads were extracted from all bam files mapped with bwa, converted to fastq files with Hydra⁴⁸ and subsequently subjected to fusion genes analysis with FusionMap⁴⁹. Chimeric transcripts that were located on the same chromosome and strand were retained for further analysis. The list of chimeric transcripts was further filtered by the number of split reads (>2) supporting the fusion. To ensure that the observed population specific fusions are not due to the lack of expression of the partner genes in the other populations, only genes with positive RPKM values observed in all populations were considered in analysis. The RPKM values were estimated with the RPKM_count.py script implemented in the RSeQC tool⁵⁰.

Finally, the novelty of the chimeric transcripts was assessed via queries against known annotation databases for read-through events such as ConjoinG⁵¹, AceView⁵², published read-through events⁵³ and literature search. Of the total 122 fusions, 75 had been identified before.

RNA editing

RNA editing is a modification of RNA transcripts that might result in alterations of coding or non-coding sequence. Here we assessed population variation of RNA editing events in RNA-seq data by calling variants at 42,039 known editing sites from the DARNED database.⁵⁴ We performed multi-sample variant calling over the 462 Geuvadis samples using SAMtools. Altogether, SAMtools called non-reference variants (i.e. at least one sample had a non-reference “genotype”) at 24,680 sites. In downstream analysis, we used only the 421 of our samples that were part of 1000 Genomes Phase 1. To reduce the number of false positive RNA editing events we applied a set of very stringent filters: (1) we required a minimum median coverage of 10 at all called sites; (2) At least 10 samples had to

have a non-reference “genotype” at each site; (3) all variants had to pass the SAMtools varFilter script; (4) the variant quality at all sites had to be above 100. Furthermore, to ensure that the observed variants are true RNA editing events and not due to unknown genetic variants, we required two things: (5) there should not be a corresponding variant in the 1000 Genomes Phase 1 data set; (6) all variants had to be located within the set of accessible regions defined by the 1000 Genomes project to ensure that a variant would be present in the genetic variant data if it was present at the DNA level.

miRNA effects on the transcriptome

miRNA family and target definition

In the analysis of association of miRNA-mRNA quantifications we used 449 samples with both miRNA and mRNA expression data. For defining miRNA-targets we used the TargetScan version 5.2 predictions.³⁰ Specifically, we downloaded the seed families of all known miRNAs conserved in vertebrates or mammals, and the corresponding conserved target sites (<http://www.targetscan.org/>). The target sites were lifted from REFSEQ annotations by mapping the 3'UTR sequences to the hg19 genome and intersecting the coordinates with our merged exon annotations (see mRNA Quantifications). The validity of the lift was confirmed at the sequence level by matching the seed sites of targets with the reverse complement of the miRNA seeds. For quantifying miRNA seed expression, we summed up read counts for all miRNAs with the same TargetScan seed sequences. E.g. the expression of the miR-141/200a seed was found by summing the read counts from hsa-miR-141-3p and hsa-miR-200a-3p. For mRNA expression data, we used the count data for the exon containing the predicted miRNA binding site. Both microRNA and mRNA expression data were corrected for hidden confounding factors with PEER and the resulting residuals were transformed to standard normal (see Normalization of quantifications). The final analysis included 100 microRNA-families and 126,698 exons.

Integrated analysis of miRNA and mRNA expression

The integrated analysis is based on the globaltest (PMID: 14693814) and is further described in (Iterson et al., Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions, submitted). Previously, it was shown that a global test-based integration model is robust and sensitive to identify sets of genes whose expression is affected by copy number⁵⁵. In this context, the globaltest was used for testing of the association of a group of genes – the predicted targets – with a microRNA expression profile. It is specifically designed for the situation of more samples than genes ($p \gg n$). Furthermore, the test overcomes the large multiple testing problem that arises when each target is tested individually for association with a microRNA expression profile. P-values for a set of target mRNAs sharing a predicted miRNA seed sequence were obtained by 100,000 permutations of the

sample labels and corrected for multiple testing using Holm's procedure. Within each set of predicted mRNA targets, P-values for individual associations between expression of predicted mRNA targets and miRNA expression levels were corrected by the Bonferroni multiple testing procedure.

A useful interpretation of the global test is as a sum of squared covariances between a set of predictors $X_{n \times p}$, and responses, $y_{n \times 1}$ (see section 5 of ⁵⁶). Consider the sample covariance, $r_{y,x}$ between a miRNA expression profile $y_{n \times 1}$ and a single target $x_{n \times 1}$ given by:

$$r_{y,x} = \frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y}_n)(x_k - \bar{x}_n) = \frac{(\mathbf{x} - \bar{\mathbf{x}}_n)^T (\mathbf{y} - \bar{\mathbf{y}}_n)}{n-1},$$

where \bar{y}_n and \bar{x}_n denote the sample means of miRNA and mRNA expression profiles, \bar{y}_n and \bar{x}_n are vectorized versions (note that $r_{y,x} = r_{x,y}$). For multiple mRNA profiles $X_{n \times 1}$ the $p \times 1$ vector of the sample covariances, $r_{y,X}$ can be expressed as:

$$\mathbf{r}_{y,X} = \frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y}_n)(X_{kj} - \bar{X}_j) = \frac{(\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{y} - \bar{\mathbf{y}})}{n-1}.$$

Note that this expression is valid even when the number of targets exceeds the number of samples $p > n$, and again $r_{y,X}^T = r_{X,y}$. Now the global test test-statistics,

$$\frac{(\mathbf{y} - \bar{\mathbf{y}}_n)^T \mathbf{X} \mathbf{X}^T (\mathbf{y} - \bar{\mathbf{y}}_n)}{(\mathbf{y} - \bar{\mathbf{y}}_n)^T (\mathbf{y} - \bar{\mathbf{y}}_n)} \propto \mathbf{r}_{y,X}^T \mathbf{r}_{y,X}$$

is proportional to the squared sample covariance.

Trans-eQTL effects of cis-mirQTLs

Variants that associate to miRNA expression levels (mirQTLs) can potentially be trans-eQTLs for the target genes of these miRNAs. This effect was sought using the European data set. The hypothesis was that a mirQTL variant should have a stronger trans-eQTL effect on the targets of the miRNA than on non-targets. This analysis is highly dependent on the accuracy of target predictions and can be conservative.

We selected all miRNA-target exon pairs based on the TargetScan predictions (see above). From these, we selected only those exons that were included in eQTL analysis (expressed in >90% samples), and only the 60 miRNAs that had a mirQTL. This left us with only 12 miRNAs. For the best-associating variant of each of the 12 mirQTLs, we collected trans association p-values (>5MB from the site) with exons that have a target site of the miRNA affected by the mirQTL (6392 variant-exon pairs in total, 125-1003 exons per mirQTL), and as a null with exons not in genes that have a target site of the miRNA with an mirQTL (4842061 variant-exon pairs in total, 81063-82518 exons per eQTL). We compared these p-value distributions for negative and positive associations separately, i.e. where the cis-mirQTL allele increasing the miRNA expression has negative or positive correlation to the exon.

Transcriptome QTL analysis

Transcriptome QTL mapping with linear regression

The details of sample sets, data filtering and normalization are discussed above. Briefly, we did transcriptome QTL mapping separately for European (n=373) and Yoruba (n=89) populations. We used genetic variants with MAF>5% in either EUR or YRI <1MB from transcription start site, with covariates of imputation status (0|1), PCs 1-3 for Europeans and PCs 1-2 for Yoruba. For the different quantitative phenotypes, we used normalized quantification units (e.g. exons) with quantification >0 in >90% of all the individuals unless mentioned otherwise (Table S3). Only autosomes were analyzed.

QTLs were mapped using a linear model implemented in Matrix eQTL⁵⁷, and FDR was estimated by permutations as follows: For exon eQTLs, we permuted the quantifications of each exon 2000 times, keeping the best p-value per exon from each round. From these data, we adjusted the FDR to 5% according to the most stringent exon of each gene, having a separate p-value threshold for each gene. For miRNAs and RNA editing sites, we ran 8000 permutations for each quantification unit, and calculated a p-value for each of them. For transcript ratio QTLs, we permuted ratios of all transcripts of randomly selected 1000 genes 3000 times and calculated a genome-wide p-value limit based on the median of the most stringent transcript per gene. For gene and repeat eQTLs, we permuted randomly selected 1000 genes and 500 repeats, and used their median as a genome-wide p-value limit.

Transcript ratio QTL effects

For the transcript ratio QTLs (trQTLs), we sought to characterize the QTL effect on transcript usage. For each trQTL gene, we identified the transcript with highest association and the transcript with most negatively correlated quantifications to this. Given the annotation of these two transcripts, AStalavista^{58,59} was used to classify the events for each trQTL.

Independence of QTLs

To estimate independent QTL signals for the same gene, we used an approach where the linear regression QTL analysis is reran using a previous association signal as a covariate – in cases where a second variant is not the same and not linked to the first one, an association signal for the gene should remain.

We applied this to estimate the number of exons with independent eQTLs from the best association for all the exons of the gene. Additionally, for 279 genes that had both a significant transcript ratio (trQTL) and a gene eQTL, we reran the eQTL analysis with the best trQTL variant as a covariate to estimate whether the trQTL signal is driving the eQTL association as well.

Null variant distribution

To compare QTL variants to a null distribution of similar variants but without regulatory association, we sampled genetic variants in cis-regions of genes expressed in our data set based on the QTL variant distributions of distance from the gene (taking upstream and downstream distance into account) and minor allele frequency. We also tried matching for the coding/noncoding status of the variants, but did not use this in the final analysis since it did not appear to have a major impact in the results.

Causal regulatory variant estimation

We estimated the probability of the best associating EUR and YRI eQTL variants being the causal regulatory variants by comparing the annotation enrichment of all loci to enrichment in those that are very likely to be causal. Specifically, we first calculated the annotation enrichment $c_{all,y}$ of the best eQTLs relative to the matched null across all eQTL loci, separately for each annotation class y . Then, we defined a subset of eQTLs where the best eQTL is likely to be causal: we binned eQTLs according to $-\log_{10}$ p-value difference between the first and the second variant (Δp), hypothesizing that for very large Δp , the first variant can be safely declared as the causal variant. To determine the Δp threshold where this point is reached, we calculated the annotation enrichment between the 1st and the 2nd variant b_y for eQTLs in each Δp bin. In both EUR and YRI, b_y saturates at $\Delta p = 1.5$, similarly in all annotation categories; thus, we reasoned that eQTLs with $\Delta p > 1.5$ can be used to estimate the amount of annotation enrichment $c_{causal,y}$ for the eQTLs where the best variant is causal. Finally, from these data, we calculated the proportion of all eQTL loci where the 1st variant is causal as $p_y = (c_{all,y} - 1) / (c_{causal,y} - 1)$.

GWAS overlap of eQTLs

We first estimated a simple overlap with exon eQTL variants and 6473 published GWAS SNPs⁶⁰ that were part of the 1000 Genomes Phase 1 data set. As a null, we collected 14 000 variants matched to the minor allele frequency spectrum of the GWAS variants. To estimate whether the GWAS overlap is particularly pronounced in the top eQTLs which would be expected if the causal variant is the same, for each GWAS variant (and the null set) that overlapped significant eQTLs, we calculated the highest eQTL rank.

The large number of significant eQTL variants and GWAS variants gives a large overlap even under the null, and with genome sequencing data we are testing a very different set of variants than the GWAS studies. This makes it challenging to identify those GWAS SNPs that are truly driven by an eQTL signal. To this end, we used a published dataset of 1213 GWAS SNPs that have been statistically shown by the RTC method to be likely to tag the same causal variant as an eQTL signal^{61,62}. From these data, we extracted 86 GWAS-eQTL loci where both the original GWAS variant and the eQTL variant were found in our data, and the recombination interval containing the original eQTL and the GWAS SNP contains a significant eQTL in our EUR data that is the strongest association for that gene. For these GWAS variants, we report the top eQTL variants in our study as putative causal GWAS variants (Table S6).

Allele-specific analysis

Allele-Specific Expression (ASE) analysis

Allele-specific expression analysis was based on binomial testing of each allelic ratio of heterozygous sites within each individual. First, we excluded sites that are susceptible to allelic mapping bias: 1) sites with 50bp mapability < 1 based on the UCSC mapability track, implying that the 50bp flanking region of the site is non-unique in the genome, and 2) simulated RNA-seq reads overlapping the site show >5% difference in the mapping of reads that carry the reference or non-reference allele. In all the analyses, we only used reads with mapping quality >150, NM>=6, and sites with base quality >10.

Next, we calculated the expected reference allele ratio for each individual by summing up reads across all sites separately for each SNP allele combination after down-sampling reads of sites in the top 25th coverage percentile in order to avoid the highest covered sites having a disproportionally large effect on the ratios. These expected REF/TOTAL ratios correct for any remaining genome-wide mapping bias as well as GC bias in each individual (Fig. S28).

Finally, for all the sites covered by >=8 reads in each individual, we calculated a binomial test of the REF/NONREF allele counts, using the expected ratio described above. Except for the NMD analysis (see below), we used only sites with >=16 reads, and sites where both alleles are observed in RNA-sequencing data in order to verify that the genotype is a true heterozygote (Fig. S28).

In many analyses, differing coverage between sites creates noise due to difference in power to call ASE. To correct for this, in many analyses we used only sites with >=30 reads (Fig. S28), and sampled all sites to exactly 30 reads. In a further analysis of ASE differences between individuals, we calculated allelic expression distances between all sample pairs as the median of absolute REF/TOTAL ratio differences of all the shared heterozygous sites between individuals after sampling the reads to 30.

Allele-Specific Transcript Structure (ASTS) analysis

Allele-specific transcript structure (ASTS) is a novel sister method of ASE, and aims at detecting differences in transcripts between the two haplotypes of an individual. As in ASE, we look at reads overlapping heterozygous coding sites, and the allele of this site in the RNAseq data tells the haplotype origin of each read fragment. The distribution of the reads to exons is then quantified.

For every sample, we first retrieved all heterozygous sites that are covered by >= 20 RNAseq reads, after mapability filter as in ASE analysis. Using the pysam package (<http://code.google.com/p/pysam/>), we scanned the bam file to extract all the reads and their mates that overlap the site, separated them to reads with REF or ALT allele, and printed out a pseudo-sam file that contains information of which SNP each read overlaps, and if it carries the REF or NONREF allele.

For this file, we ran our standard exon quantification, and calculated the number of REF and ALT read overlaps in all the exons. We kept only exons with >=10 reads of each allele, and required a total of >=20 REF and NONREF reads in

the remaining exons. We used Fisher test to estimate whether the read counts in exons are different for REF and NONREF reads. For each site, we calculated a quantitative measure analogous to ASE allelic ratio (maximum imbalance across all exons of a site compared to the total REF/NONREF ratio).

Loss-of-function analysis

Nonsense-mediated decay

To estimate the signal of nonsense-mediated decay (NMD) in premature stop variants, we quantified ASE using allelic read count data from individuals who are heterozygous for a premature stop variant, compared to other individuals where we have ASE data from the same gene as the ASE variant. We applied an EM algorithm to fit a mixture of binomial distributions where number of components, k , was set to 2, and no prior information was given for the binomial distribution parameters. The EM algorithm was run until $\epsilon < 1e-8$; final number of iterations = 20. This was ran for all variants and for rare (minor allele counts = 1-10) premature stop variants.

Splice scores

For the 1000 Genomes Phase 1 SNPs and indels that modify the splice site motif, we computed log-odd scores of variant effect in splice motifs employing the 1st order Markov Models for splice donor and acceptor sites of human U2-dependent introns from the gene prediction program GeneID⁶⁴. The scoring has been applied to the ~478,000 splice sites currently included in the Gencode v12 reference annotation.

Splice site variants have been inferred from the 1000 relevant for the Markov model.

The Geuvadis Data Browser

For the visualisation of RNA-sequencing analysis we created the Geuvadis Data Browser (www.ebi.ac.uk/Tools/geuvadis-das). It is powered by the Genoverse browsing engine running HTML5 and Javascript and co-developed by the Ensembl and DECIPHER projects. The back-end for the browser is the EBI data sources providing the Geuvadis analysis data in real-time.

The Geuvadis RNA-sequencing analysis results consist of following tracks:

- EUR and YRI exon eQTLs,
- EUR and YRI exon quantifications,
- EUR and YRI mirQTLs,
- EUR and YRI mirRNA quantifications,

Quantification tracks show the population minimum, average and maximum values of raw counts normalised by library size and element lengths, very similar to FPKM normalization. By clicking on the element of interest it is possible view information about each element: description, scoring information and links to other relevant data sources, for instance Ensembl Genome Browser.

QTL tracks show SNPs and indels associated with functional effects. In a similar way a click on an element of interest will provide additional information including all linked effect elements associated with eQTL along with related p-values.

Tracks at the top of the Geuvadis Data Browser provide gene and transcript element annotations. These tracks are based on Ensembl latest release of human genome GRCh37 and are given for the reference purposes. It is possible to search for genes, variants or locations.

Individual bam files can be viewed at the read level in ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1>), by accessing the link to the Ensembl Genome Browser under the detailed sample information and linked data view.

References to Supplementary Methods

- 28 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:btp352 [pii]
10.1093/bioinformatics/btp352 (2009).
- 29 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-825, doi:nbt.1662 [pii]
10.1038/nbt.1662 (2010).
- 30 Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20, doi:S0092867404012607 [pii]
10.1016/j.cell.2004.12.035 (2005).
- 31 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157, doi:gkq1027 [pii]
10.1093/nar/gkq1027 (2011).
- 32 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).
- 33 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:ng1847 [pii]
10.1038/ng1847 (2006).
- 34 Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, doi:nmeth.2221 [pii]
10.1038/nmeth.2221 (2012).
- 35 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:btp324 [pii]
10.1093/bioinformatics/btp324 (2009).

- 36 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:22/9/1760 [pii] 10.1101/gr.135350.111 (2012).
- 37 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:nature08903 [pii] 10.1038/nature08903 (2010).
- 38 Shapiro, I. M. *et al.* An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet* **7**, e1002218, doi:10.1371/journal.pgen.1002218 PGENETICS-D-10-00244 [pii] (2011).
- 39 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:nature07509 [pii] 10.1038/nature07509 (2008).
- 40 Buermans, H. P., Ariyurek, Y., van Ommen, G., den Dunnen, J. T. & t Hoen, P. A. New methods for next generation sequencing based microRNA expression profiling. *BMC Genomics* **11**, 716, doi:1471-2164-11-716 [pii] 10.1186/1471-2164-11-716 (2010).
- 41 Pantano, L., Estivill, X. & Marti, E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res* **38**, e34, doi:gkp1127 [pii] 10.1093/nar/gkp1127 (2010).
- 42 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467, doi:84979 [pii] 10.1159/000084979 (2005).
- 43 Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* **34**, D158-162, doi:34/suppl_1/D158 [pii] 10.1093/nar/gkj002 (2006).
- 44 Berninger, P., Gaidatzis, D., van Nimwegen, E. & Zavolan, M. Computational analysis of small RNA cloning data. *Methods* **44**, 13-21, doi:S1046-2023(07)00176-4 [pii] 10.1016/j.ymeth.2007.10.002 (2008).
- 45 Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**, e1000770, doi:10.1371/journal.pcbi.1000770 (2010).
- 46 Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigo, R. Estimation of alternative splicing variability in human populations. *Genome Res* **22**, 528-538, doi:gr.121947.111 [pii] 10.1101/gr.121947.111 (2012).
- 47 tweedEseq: RNA-seq data analysis using the Poisson-Tweedie family of distributions. v.1.0.14.
- 48 Quinlan, A. R. *et al.* Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**, 623-635, doi:gr.102970.109 [pii] 10.1101/gr.102970.109 (2010).

- 49 Ge, H. *et al.* FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27**, 1922-1928, doi:btr310 [pii]
10.1093/bioinformatics/btr310 (2011).
- 50 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-2185, doi:bts356 [pii]
10.1093/bioinformatics/bts356 (2012).
- 51 Prakash, T. *et al.* Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. *PLoS One* **5**, e13284, doi:10.1371/journal.pone.0013284 (2010).
- 52 Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7 Suppl 1**, S12 11-14, doi:gb-2006-7-s1-s12 [pii]
10.1186/gb-2006-7-s1-s12 (2006).
- 53 Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med Genomics* **4**, 11, doi:1755-8794-4-11 [pii]
10.1186/1755-8794-4-11 (2011).
- 54 Kiran, A. & Baranov, P. V. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26**, 1772-1776, doi:btq285 [pii]
10.1093/bioinformatics/btq285 (2010).
- 55 Menezes, R. X., Boetzer, M., Sieswerda, M., van Ommen, G. J. & Boer, J. M. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics* **10**, 203, doi:1471-2105-10-203 [pii]
10.1186/1471-2105-10-203 (2009).
- 56 Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93-99 (2004).
- 57 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358, doi:bts163 [pii]
10.1093/bioinformatics/bts163 (2012).
- 58 Sammeth, M., Foissac, S. & Guigo, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol* **4**, e1000147, doi:10.1371/journal.pcbi.1000147 (2008).
- 59 Foissac, S. & Sammeth, M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res* **35**, W297-299, doi:gkm311 [pii]
10.1093/nar/gkm311 (2007).
- 60 Hindorff, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies (2010).
- 61 Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-1089, doi:ng.2394 [pii]
10.1038/ng.2394 (2012).
- 62 Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895, doi:10.1371/journal.pgen.1000895 (2010).

- 63 Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* **23**, 198-199, doi:S0968-0004(98)01208-0 [pii] (1998).
- 64 Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 3, doi:10.1002/0471250953.bi0403s18 (2007).

Supplementary Tables

			QC-passed			
			mRNA		miRNA	
Pop	Full name	Total sequenced samples	Total	1000G Phase1	Total	1000G Phase1
CEU	Utah residents (CEPH) with Northern and Western European ancestry	92	91	78	87	74
FIN	Finnish from Finland	95	95	89	93	87
GBR	British from England and Scotland	96	94	85	94	84
TSI	Toscani in Italia	93	93	92	89	88
YRI	Yoruba in Ibadan, Nigeria	89	89	77	89	77
TOT	Total	465	462	421	452	410

Table S1. Samples

Numbers of sequenced individuals. Replicate samples are not included in the counts.

Coding annotation (nonredundant hierarchy)	Variants
SPLICE_DONOR_VARIANT	4036
SPLICE_ACCEPTOR_VARIANT	2977
STOP_GAINED	6483
FRAMESHIFT_VARIANT	1186
STOP_LOST	581
INITIATOR_CODON_CHANGE	1034
INFRAME_CODON_GAIN	193
INFRAME_CODON_LOSS	531
NON_SYNONYMOUS_CODON	305959
SPLICE_REGION_VARIANT	53901
INCOMPLETE_TERMINAL_CODON_VARIANT	29
SYNONYMOUS_CODON	197584
STOP_RETAINED_VARIANT	253
CODING_SEQUENCE_VARIANT	31
COMPLEX_CHANGE_IN_TRANSCRIPT	97
MATURE_MIRNA_VARIANT	432
5_PRIME_UTR_VARIANT	101725
3_PRIME_UTR_VARIANT	381972
INTRON_VARIANT	19734371
NC_TRANSCRIPT_VARIANT	190673
Noncoding annotation (redundant)	Variants
MIRNA_TARGET	3324
TFMOTIF	50282
REG_FEATURE	7325520
ACTIVE_CHROM	38137117
MIRNA_MATURE	652
MIRNA_PRECURSOR	1290
NOVEL_SPLICE	431
No annotation	1027762

Table S2. Variant annotations

Numbers of variants in annotation categories.

	in >50% of samples	In QTL analysis
Genes	16084	12981
Transcripts	67603	13704
Exons	146498	122893
Splice junctions	12805	NA
Transcribed repeats	47409	43875
Chimeric transcripts	5	NA
RNA edited sites	99	99
miRNAs	715	644

Table S3. Quantifications

Numbers of quantified transcriptome features. Gene, transcript, exon and annotated splice junction counts are from protein-coding and lincRNA genes. All eQTL counts are for autosomal genes, with a filter of quantification in >90% of samples for genes, exons, transcripts, and transcribed repetitive elements, and >50% for miRNAs and RNA edited sites.

Tables S4-S6 are available as separate files:

Table S4. Chimeric transcripts (legend)

Detected chimeric transcripts.

Table S5. Associated miRNA-mRNA pairs (legend)

List of 36 significant ($P < 0.001$, Holm) miRNA families and their associated mRNA targets ($P < 0.05$, Bonferroni). The column descriptions are:

- Exon (exon identifier consisting of Ensembl gene id, chrom location, start and end exon containing the predicted microRNA binding site; exons are unions of all overlapping exons of the same gene)
- microRNA family: family of microRNAs with identical seed-regions
- P-value (of set): P-value indicating the strength of association of the microRNA expression profile with the set of predicted targets
- P-value (target): P-value indicating the target's individual contribution to the overall strength of association to the set
- Association: '0' indicates negative association of the microRNA expression profile with the predicted targets and '1' positive association
- Entrez Gene: Entrez gene identifier
- Gene Symbol: HGNC gene symbol

Table S6. Predicted causal GWAS variants (legend)

GWAS variants that have a signal of a shared causal variant with an eQTL (see Supplementary Methods), and the eQTL p-values of the top eQTL variants and the GWAS SNP.